Quasi-Experimental Evaluation

Evaluation and Data Development
Strategic Policy
Human Resources Development Canada

January 1998

SP-AH053E-01-98
également disponible en français

Acknowledgement

Evaluation Services of HRDC has traditionally used a quasi-experimental research design (comparison group methodology) in estimating the impacts of an intervention. In order to ensure the suitability of the comparison group approach in subsequent evaluations of programs and services delivered under EI Part II, the attached study "Quasi-Experimental Evaluation" was carried out.

We would like to thank Dr. Robert Power of Power Analysis Inc. of London, Ontario and Dr. W. Craig Riddell, Department of Economics, University of British Columbia for researching and analyzing the evidence for and against comparison group design and for the preparation of the report on findings from a through review of the literature.

Series

Human Resources Development Canada (HRDC) has a long tradition in program evaluation. Over the years, the Evaluation and Data Development Branch (EDD) of HRDC has conducted numerous evaluations on a wide range of departmental programs. The expertise and experience acquired have permitted EDD to hone its evaluation techniques and approaches.

More recently, however, changes in the delivery nature of some HRDC programs have affected the way evaluation is conducted. Now, HRDC regional offices are more directly and actively involved in program evaluation.

In response to these changes EDD is proud to initiate a new series of publications entitled "Evaluation Tool Kit". The objective of the series is to build evaluation capacity and knowledge throughout HRDC and its partner organisations by providing pertinent information about designing, planning and conducting an evaluation. Existing information about particular topics will be synthesised and made readily accessible in short, precise reports which are adapted to the evaluation of human resources development programs.

The evaluation field is one of growing interest and progress. We invite you to provide us with comments you may have on this series or on a specific study.

The first study of the new series focuses on the quasi-experimental approach in evaluation.

Table of Contents

1.0 Definition of Evaluation **3**
2.0 Two Main Types of Evaluation **3**
3.0 Summative Evaluation Designs **4**
4.0 Types of Designs **4**
5.0 Examples of Selection Bias **9**
6.0 Procedures for Addressing Selection Bias **11**
6.1 Two Step Adjustment Procedures 12
6.2 Instrumental Variable Methods 13
6.3 Longitudinal Methods 15
6.4 Specification Tests of Alternative Models 16
6.5 Determinants of Program Participation 17
7.0 Selecting Outcome Measures **17**
8.0 Selecting a Comparison Group **18**
9.0 Applying the Lessons: Choosing Comparison Groups for EBSM **19**
10.0 Conclusion **20**

Appendix A: True Experiments Versus Quasi-Experiments **21**
Appendix B: Brief Mathematical Representation of Quasi-Experiments **25**
Appendix C: Dealing With Selection Bias **27**
Appendix D: Determining Appropriate Sample Size **29**

References **31**

Quasi-Experimental Evaluation

Program manager: Why do we need to evaluate our program? We have a good handle on what's going on with our program and our clients, and we know we are very successful.

Evaluator: Because you never know if it was your program or something else that produced the success you are claiming.

Program manager: Of course we know it's our program. What else would cause all these people to find jobs so quickly?

Evaluator: Maybe the rapid economic expansion that we are now enjoying.

Program manager: That's nonsense. Anyway, we know we have to have our program evaluated. But why do we need to go through the trouble of finding a comparison group to do an evaluation?

Evaluator: Let's say that six months after finishing your training program, 70% of the trainees are working. Would you consider that proof your program is a success?

Program manager: I'd say so, yes. We'd like to do better than 70% —in fact, we believe we are doing better than that — but I'd say that having 70% of our trainees working would make our program look very good, especially given the barriers many of our clients face when they come to us.

Evaluator: What proportion of these individuals might be working now if they hadn't gone through your training program?

Program manager: I'm not sure, but it wouldn't be as high as 70%, I can tell you that.

Evaluator: Well you don't really know that though. For all you know, 80% might be working now if they hadn't taken training.

Program manager: No way. You don't know how many obstacles our clients face when they come to us. We are providing a valuable service and are really helping our clients.

Evaluator: That may be so, but you haven't proven it. And the sponsors of the program need to know with certainty how successful you have been. They need to know they are getting bang for their buck.

Program manager: They are, I assure you.

Evaluator: Okay, let's say you are doing some good, that individuals who go through your program are indeed more likely to find a job than if they hadn't been trained. How much of an effect are you having? Would half of them have found jobs anyway? One-third? Two-thirds? You can't know that unless you do an evaluation that includes a comparable group of people who haven't taken your training?

Program manager: Even if half of them would have found a job without the training, isn't raising that proportion to 70% worth it?

Evaluator: I don't know. What did it cost for that incremental 20%? And how long will the effects of training last?

Program manager: Our program is well worth the money . . .

This contrived conversation illustrates well the different perceptions of program managers and evaluators when it comes to assessing the merits of a program. Managers live and work with the program every day: they care about their program and work hard to make it a success; they strongly believe they are doing a good job; and they understandably resent any implication that they are not.

Evaluators usually have no connection with the program, and, more importantly, no stake in its survival (which sometimes leads to an underestimation of the threat that evaluations can impose on program management and staff). They know that managers are heavily invested in their program and that a manager's assessment of the program — even one aided by reliable monitoring data — will not be accepted by program sponsors as a valid test of whether the program is meeting its objectives and is worth what it costs. And they know that many different factors that are unrelated to the design of the program can affect the outcomes of any social program, and can easily lead to unwarranted conclusions about the program.

This paper is based on the premise that only a good program evaluation can produce convincing evidence of a program's effectiveness in reaching its objectives, and that only an evaluation that includes a control or comparison group can provide estimates of program impact uncorrupted by the influence of other factors that also may affect outcomes. It summarizes the basics of evaluation research, focusing on what is usually the best practical approach — the "quasi-experimental design." It is written in non-technical language for managers who have little or no exposure to the field of evaluation, but includes more advanced treatments in appendices for those interested in the more technical aspects.

The paper begins with a brief introduction to evaluation, including a broad definition, and a summary of the two basic types of evaluation. It then moves to a discussion of evaluation design, starting with the reasons evaluators need to worry about appropriate designs. Common evaluation designs are introduced, with the pros and cons of each discussed. Designs without comparison or control groups are shown to fall well short of ideal in terms of ruling out extraneous

causes. Since any good evaluation is fundamentally a comparison between what happened to program clients and what would have happened had they not been in the program, one-group designs virtually preclude any serious summative evaluation.

Next, the report explains details of quasi-experimental evaluation design — its theoretical underpinnings and practical considerations. The concept of selection bias is explained and econometric procedures to control for it are introduced. Confidence in econometric estimates will be explored.

How to select comparison groups is the next consideration. The different techniques for matching are summarized, with an assessment of their relative strengths and weaknesses. Potential variables to use in conducting the matching are reviewed at this stage. Finally, with the general topic of quasi-experimental evaluation well explicated, we then proceed to apply the lessons to the more specific tasks of determining the best sampling techniques for drawing regional Employment Benefits and Support Measures (EBSM) comparison groups, and the best variables to use in the sampling.

1.0 Definition of Evaluation

There are a lot of different definitions of evaluation. Here is one of the best, because it touches on the most important aspects of evaluation:

> "Evaluation is a collection of methods, skills and sensitivities necessary to determine whether a human service is needed and likely to be used, whether it is conducted as planned, and whether the human service actually does help people" (Posavac and Carey, 1980, p.6).

This definition encompasses the two main types of evaluation: process and summative.

2.0 Two Main Types of Evaluation

Although the literature includes over a hundred different kinds of evaluation (see Patton, 1982), the vast majority boil down to two types: those that aim to determine if the program has been implemented as planned, and those that measure its success in achieving its objectives (i.e., its impact). The label most often associated with the first type is "process evaluation," although it is sometimes called formative evaluation. The latter type is known as "summative evaluation," also known as impact, outcome, or effectiveness evaluation.

Process Evaluation — *How is the program operating and how can it be made better?*
Process evaluations are directed at three key questions: (1) the extent to which a program is reaching the appropriate target population; (2) whether or not its service delivery is consistent with program design; and (3) what resources are being expended (Rossi and Freeman, 1993).

The main objective is to provide feedback to managers on whether the program is being carried out as planned and in an efficient manner. Guidance should be provided for modifying

the program to help ensure it meets its objectives. With this information, the program can be modified so it is carried out as planned, or the plan itself can be modified if it is found lacking.

Summative Evaluation — *Does the program achieve its objectives?* The purpose of summative evaluations is to assess the impact of the program; that is, ascertain the extent to which the program meets its objectives, and the needs of its target group. As well, it should provide advice for modifying the program so that it will better serve the needs of its clients and become more cost-effective (Stufflebeam and Shinkfield, 1985).

The remainder of this report will focus on the summative side, in particular summative evaluations with quasi-experimental designs. Before plunging into the arcane field of quasi-experimental design however, the paper sets a proper context through a discussion of the main types of summative evaluation designs.

3.0 Summative Evaluation Designs

There is no single correct evaluation design for impact evaluations. The goal is to come up with the best design possible under the circumstances. Almost all designs represent a compromise dictated by many practical considerations such as how much money and time are available, what the client considers compelling, how much a design might interfere with the normal operation of the program, and so on.

In deciding on the design, the credo is to maximize the credibility and usefulness of the findings. The evaluator must anticipate the kind of arguments that will be used to dismiss the findings.

In summative evaluations the key concern is to be able to ascribe the outcome to the program as opposed to innumerable other possible causes. In the vernacular of evaluators, this is known as "***internal validity.***" Every good summative evaluation is designed to minimize "threats to internal validity"; that is, they are designed to isolate the impact of the program from the impact of other potential causes. Campbell and Stanley (1971) identified seven threats to internal validity (Exhibit 1).

4.0 Types of Designs

There are dozens of possible designs to determine program impact. This section will summarize the most popular ones.

Because designs without comparison groups are very deficient in terms of internal validity — that is, they cannot normally rule out alternative explanations for program outcomes observed — ***all good summative evaluations include a comparison group.*** Nevertheless, single group designs are very common.

Exhibit 1 - Threats to Internal Validity

Threats due to real changes in the environment or in participants:

History: Changes in the environment that occur at the same time as the program and will change the behaviour of participants (e.g., a recession might make a good program look bad).

Maturation: Changes within individuals participating in the program resulting from natural biological or psychological development.

Threats due to participants not being representative of the population:

Selection: Occurs when assignment to participant or non-participant groups yield groups with different characteristics. Pre-program differences may be confused with program effect.

Mortality: Participants dropping out of the program. Drop-outs may be different from those who stay.

Statistical Regression: The tendency for those scoring extremely high or low on a selection measure to be less extreme during the next test. For example, if only those who scored worst on a reading test are included in the literacy program, they might be bound to do better on the next test regardless of the program just because the odds of doing as poorly next time are low.

Threats generated by evaluators:

Testing: Effects of taking a pretest on subsequent post-tests. People might do better on the second test simply because they have already taken it [1].

Instrumentation: Changes in the observers, scores, or the measuring instrument used from one time to the next.

Single Group Designs

The simplest but least satisfactory evaluation design is the posttest only design, symbolized, X O (where X is the program intervention — e.g., the training course — and O is a post-program observation such as annual earnings). Here, participants, having completed the program of interest, are surveyed to find out how well they are doing with respect to the behaviours or attitudes at issue. *This design cannot be used to credibly attribute any effects to the program*, for there is no objective basis to suppose that the program caused any changes. Indeed, because there is no information on the pre-program level of the variable(s) of interest, this design yields no information on change.

---

[1] Also, taking a pretest may sensitize participants to a program. Participants may perform better simply because they know they are being tested — the "Hawthorne effect."

Whenever a program is supposed to bring about a change, before-and-after measures are a necessity with one-group designs. The simplest, the *pretest-posttest* design symbolized as O_1 X O_2 (where X again denotes the intervention and O1 and O2 denote pre- and post-program outcome measures) requires a pretest of some sort before the program takes place (a reading test, for example), and a posttest after the program. This design is subject to most threats to internal validity. Most seriously, participants might have changed or some extraneous event may have brought about any observed difference between O_1 and O_2, so *no change can credibly be ascribed to the program.* For example, if an evaluation of this design showed that mean post-program earnings were lower than mean pre-program earnings, this should not be construed as proof that the program was defective: a recession may have caused the earnings drop.

Time series designs involve collecting data repeatedly about participants' situation at several times. Symbolically:

$$O_1 O_2 O_3 X O_4 \ldots O_n$$

This design can be used to rule out (or at least quantify) regression and maturity as threats to internal validity. That is, any personal trends in the absence of the program can be accounted for and controlled. Advanced statistical procedures are required to isolate the effect of the program. History remains the key threat. Although supplementary data on the environment can help rule out events that can be identified, it is extremely difficult to identify — let alone quantify — all possible events that could have brought about the outcome observed. Consider the following example. Say Province A implemented a large-scale training program for its social assistance clients and observed the social assistance caseload statistics for several months before and after the intervention to see if the training program was lowering dependence on social assistance. But, at around the same time, Province B instituted its own policy change: a cutback in social assistance benefits for employable clients. That could precipitate an inflow of social assistance clients from Province B to Province A. Unless Province A knew about the policy change in Province B and took steps to measure its impact, the time-series evaluation could underestimate any positive impact of the training program.

In sum, one-group designs virtually preclude any serious summative evaluation. They are notoriously weak and easily dismissed, because it is generally impossible to rule out potential alternative explanations, especially key events that occurred while the treatment group was participating in the program (e.g., a recession). Two-group designs are required for sound evaluation.

Two Group Designs

A valid determination of impact requires comparing outcomes of a group of individuals who have participated in the program (treatment group) with an equivalent group of people who have not participated (control or comparison group). In theory, the best way to do this is by means of a *randomized experiment*, where individuals are assigned at random to the treatment or control group (Rossi and Freeman, 1993). Outcomes measures, chosen on the basis of

program objectives, are observed at some interval after the intervention ends, with any differences between groups attributable to the program: that is, the program can be said to have caused the observed differences. The design is symbolized as follows:

X O [participants]
O [non-participants]

Since randomization should remove — at least on average — any systematic differences between the groups, no pretest is needed. The impacts of the treatment can be measured simply by comparing the mean outcomes for treatment and control groups, with chance differences largely accounted for through standard statistical techniques (Greenberg and Wiseman, 1992).

The primary threat to this design — assuming the randomization process was carried out correctly — is non-random mortality or attrition (i.e., members of the participant group drop-out before completing the program, or members of participant and control groups cannot be located for follow-up purposes for reasons that are not random and therefore potentially systematically related to the impacts of the program). For this reason, a pretest is often given to both treatment and control groups (Mark and Cook, 1984, hold this is "essential"), so the effects of discontinuation from the program can be quantified and accounted for in the analysis.

Although experimental designs are as close to ideal as possible in theory, they are seldom practical. By far the most common constraints are program staff who refuse to comply because they consider randomized selection unethical or unacceptable, and evaluation timing: most often the evaluator enters the scene long after random assignment should have taken place. There are other problems as well. Most seriously, experimental methods are normally confined to a determination of the mean impact of the program; they cannot answer many other key policy questions, including the median impact of the program and the proportion of participants with a positive (or negative) impact from the program (Heckman and Smith, 1995)[2].

Given these constraints, quasi-experimental (non-experimental) models are frequently the only satisfactory way to proceed. There are different *quasi-experimental models*, but the most common and robust method involves constructing a comparison group of individuals who are comparable to participants. This can be done: by statistically controlling for differences between groups during data analysis, by matching participants and non-participants according to key traits (such as age, sex and education) believed to influence the outcomes of interest; or both[3].

---

[2] Appendix A briefly explores the on-going debate on the choice between experimental and quasi-experimentalmethods.

[3] Rubin (1979) showed that the use of both techniques — matching and statistical adjustment — was better than use of either technique alone (as reported in Dickinson et al, 1987).

The idea is to approximate random assignment as closely as possible by attempting to minimize or control for differences between the groups. Symbolically[4]:

$$O_1 X O_2 \quad \text{[participants]}$$
$$\overline{O_1 O_2} \quad \text{[non-participants]}$$

Here X is the program intervention, O_1 is a pre-program observation, and O_2 is a post-program observation. For instance, O_1 could be annual earnings in 1995, O_2 could be annual earnings in 1997, and X could be a 1996 training program.

Under a longitudinal quasi-experimental approach, for example, the evaluator compares the outcomes of two groups: program participants (the "treatment group") and non-participants (the "comparison group")[5]. "Outcomes," which relate to the objectives of the training program — finding a job, for example — are usually determined via a follow-up survey, conducted months or even years after program exit. The post-program outcome for each group is compared to pre-program status to determine if there has been any change, on average, within each group. Then, standard statistical procedures determine whether the change differs significantly between groups.

For example, say a baseline survey (see below) showed that half the participants and half the non-participants were working one year prior to applying for EI; and a follow-up survey found that 70% of trainees were working one year after the training, versus 60% of the comparison group one year after leaving EI. The increment for trainees is thus 20 percentage points, versus 10 for non-trainees. Simple statistical tests would determine if this difference is significant.

But, a finding of a statistically significant difference between groups does not necessarily imply that the difference was due to the program. *The analyst must demonstrate that the difference is attributable to the program.* That is, threats to internal validity must be ruled out. Unfortunately, the empirical evidence shows that participants are likely to be different from non-participants in ways that affect the outcome variables. Selection into most programs is non-random: those who volunteer to participate may be more motivated than those who do not, for example; and program administrators more often than not select those they feel will have the best chance of succeeding (i.e., the most talented), or conversely, select those most in need of the treatment.

Regardless of its source, *selection bias* affects the comparability of treatment and comparison groups. As long as all differences between the groups being compared are observable (e.g., personal traits), selection bias will not be a problem because statistical methods such as multiple

---

[4] Note that pre-program information is not strictly necessary for some comparison group designs — see section below on addressing selection bias — but it is always desirable.

[5] Appendix B presents an elementary mathematical representation of the quasi-experimental method.

regression analysis can control for the differences. Researchers do their utmost to match individuals in treatment and control samples to ensure observed characteristics are very similar, but they seldom know why a person is participating in a program. If any unknown (hence uncontrolled) feature of the person or program influenced the decision to participate, then the selection is non-random and differences between participants and non-participants may be incorrectly ascribed to the program.

No statistical method is likely to completely resolve the selection bias problem. Since it is impossible to anticipate all the factors that went into the decision to participate, the surveys and protocols cannot be designed to gather all relevant information. Quasi-experiments require analysis techniques that are much more complicated than those for true experiments. High-level statistics — "econometric models" — are required to deal with the differences between groups and isolate the effect of the program.

5.0 Examples of Selection Bias

A number of procedures are available for dealing with the potential problem of selection bias. In order to describe these methods it is helpful to use some concrete examples of the forms that selection bias frequently takes in evaluation research on education/training programs.

Example 1: Coop versus non-coop education

Cooperative education programs combine periods of formal education with periods of work experience in a systematic manner, while the more common non-coop educational programs provide formal education without organized periods of work experience. Many observers feel that the combination of education and work experience is likely to enhance the employability and earnings of coop graduates. Does empirical evidence support this belief? A simple way of answering this question would be to compare the post-graduation employment and earnings experiences of a sample of coop and non-coop graduates. Would this be convincing evidence of the impact of coop programs on employability and earnings? The answer in general is no.

The reason is that there may be reasons why those completing coop programs may have different levels of employability and earnings than those completing non-coop programs that are independent of the educational programs themselves. For example, if there are a limited number of coop programs (as is generally the case), these programs may be able to admit more qualified students, on average, than comparable non-coop programs. Similarly, the students who apply for coop programs may be more career-oriented, on average, than those applying to non-coop programs. Both these reasons would lead to the coop graduates having higher levels of employability and earnings even in the absence of completing a coop program. That is, if the coop graduates had instead completed non-coop programs, their employability and earnings would have been higher than the non-coop graduates. To some extent, higher levels of employability and earnings of the coop graduates are due to their being better, more qualified students and to their being more career-oriented than their non-coop counterparts.

Of course, it is also possible that the coop programs have a positive impact on the employability and earnings of graduates. If so, the total observed difference between coop and non-coop graduates consists of two components: one because coop programs attract more qualified and more career-oriented students (the selection effect), and one due to the impact of the program (the program impact effect).

This is the problem of potential selection bias. Coop programs are selecting the more qualified and career-oriented students who would have had higher employability even in the absence of the program. In these circumstances, a simple comparison of coop and non-coop graduates would over-estimate the true impact of the program. However, as discussed below, selection bias may be either positive or negative; that is, without accounting for selection bias, the estimated impact may be either above or below the true impact.

Example 2: Private schools versus public schools

Many of the same issues arise if one wishes to compare the outcomes (such as average grades in standardized examinations, high school completion rates, or post-secondary success) of private and public schools. Private schools may be able to select more qualified students, on average, than their public school counterparts. Similarly, the average student attending a private school may be more likely to have other attributes (such as having parents who place a higher value on education) than the average student attending public school.

For these reasons, a simple comparison of student outcomes in private and public schools is unlikely to give unbiased estimates of the impact of private schools on these outcomes. Some of the observed differences arise because private and public schools select students who differ systematically on such attributes as qualifications, parental wealth and family attitudes towards education.

Example 3: Impact of government-sponsored training programs

As will now be clear, similar issues of selection arise in the context of assessing the impacts of training programs. Those who undertake training are likely to differ systematically from those who do not take training. This could be due to differences between the trainees and non-trainees themselves; for example, those who are more educated or more labour market oriented may be more likely to apply for training. Alternatively, these differences could arise because of selection by program administrators, who may be more likely to enroll those who are most likely to benefit from training.

This example also helps make clear why selection bias may be either positive or negative. Suppose a program is designed to focus on the most disadvantaged among a particular population. In these circumstances, a simple comparison of trainees and non-trainees is likely to under-estimate the true impact of the program.

These three examples illustrate the point that virtually any program evaluation is likely to face

some potential selection bias. The reason is that participation in virtually all programs or interventions involve choices on the part of participants and non-participants and on the part of those administering the program. Because of such choices, the pools of participants and non-participants are likely to differ in systematic (or non-random) ways. If such differences between participants and non-participants are also related to the program outcomes, simple comparisons of participants and non-participants will give biased estimates of program impact. To deal with the very pervasive problem, a number of statistical methods have been developed.

6.0 Procedures for Addressing Selection Bias

Before describing the various procedures that have been developed and extensively used to deal with potential selection bias, two preliminary observations are noted. First, as the above examples make clear, selection into the program may be based on observable or unobservable factors. For example, in the case of comparing coop and non-coop programs, the qualifications of the coop and non-coop students may be observed (e.g. their high school grades) but the degree of career-orientation of the students may not. What factors are observed and what factors are unobserved will depend on the richness of the available data.

Controlling for selection into the program which took place on the basis of observable factors is straightforward. Thus the richer the available data (and thus the fewer the number of unobservable factors), the smaller will be the magnitude of selection bias due to unobserved factors.

The second observation is that although there are some unobserved factors which influence selection into the program (which will almost always be the case), this does not necessarily imply that simple comparisons of participants and non-participants will be subject to selection bias. Selection bias arises when the unobserved factors which influence participation/non-participation in the program also influence the impacts of the program.

To make this point clear, consider an extreme example. Suppose we wish to compare public and private schools in terms of their educational outcomes such as student performance on standardized tests. Suppose that, on average, private schools enroll more students with odd numbered birthdays than public schools. To the researcher, the attribute "having an odd/ even birthdate" is unobserved. This is a case of non-random selection; if students were randomly selected into public and private schools then the proportion of students with odd numbered birthdays would be approximately equal in the two school types. However, as long as having an odd numbered birthday does not influence the outcomes of interest (student performance on standardized tests), this non-random selection will not bias a simple comparison of student achievement in public and private schools.

For these reasons, methods for dealing with selection bias focus on the potential problems associated with factors which influence selection into the program which are (i) unobserved by the researcher/evaluator, and (ii) correlated with the program outcomes of interest. Unfortunately, in many evaluation studies there are a number of such factors meeting both

these conditions. For this reason, it is strongly advisable in virtually any evaluation study to address potential selection bias.

By their very nature, the factors that might give rise to selection bias are unobserved. In some cases, the methods described below will lead the evaluator to conclude that such potential sources of selection bias are not quantitatively important, and therefore do not lead to bias. This could be because unobserved factors leading to non-random selection into the program are not quantitatively significant in this particular program, or it could be because such factors are quantitatively significant but are not correlated with the program outcomes of interest. In other cases, the methods described below will lead the evaluator to conclude that selection bias is quantitatively important. In these cases, the methods also provide an estimate of the magnitude of the bias so that an estimate of the true impact(s) of the program on outcomes can be derived.

6.1 Two Step Adjustment Procedures

Two step (or two stage) procedures for addressing selection bias were developed by James Heckman and others in the late 1970s, and have become the most commonly used methods. In the first stage, the probability of participation in the program is analyzed. This analysis usually consists of a single equation model in which the dependent variable is the probability of participating in the program (an indicator variable which equals unity for program participants and zero for non-participants) and the independent variables are various factors that are believed to influence program participation/non-participation. The main purpose of the first stage is to obtain a correction factor (called the "inverse Mills ratio") which is used in the second stage to take account of possible selection bias. As well, the estimates obtained in this first stage may be of interest in themselves in that they provide insight into the importance of the various factors that influence participation/non-participation in the program.

The second stage involves estimating program impact using a specified model (equation). The model includes:

- a "dependent variable," which is the outcome the training program is supposed to affect, say earnings;
- several "independent" or explanatory variables, which are observed factors presumed to influence the outcome (e.g., age, sex, education);
- the "selection bias correction" variable (or inverse Mills ratio) obtained in the first stage; an indicator variable for participation/non-participation in the program; and
- a random error term to account for unobserved forces that may affect the outcome measure.

The model in words:

> Earnings = the effect of various observed factors + the effect of selection bias + the effect of the program + random error

(see Appendix C for the mathematical equation and further explanation)

Thus, the model isolates the impact of the program from other potential influences. *If the model is properly specified*, the addition of the "selection bias correction" variable removes this potential bias, thus giving unbiased estimates of program impact. We return below to the important issue of how to determine whether the model is properly specified.

A useful way to interpret this two step procedure is as follows. It is well known that omitting an important variable from a model will result in biased estimates of the coefficients on the variables included in the model. In the absence of a method for accounting for selection into the program, the estimation of the outcome equation omits an important factor — the determinants of program participation. The "selection bias correction" term obtained in the first stage provides an estimate of this factor, which is why including this term results in unbiased estimates (providing the model is properly specified).

One final observation relating to these two step procedures is in order. It is important to have one or more variables that influence selection into the program (i.e., that enter the first stage equation) but which do not influence the outcome(s) of the program (i.e., do not enter the second stage equation). Such variable(s) allow the separate identification of participation in the program and the outcomes or impacts of the program. Apart from the importance of having one or more such "identifying variables," the first stage participation equation and the second stage outcome equation may have many variables in common. The importance of these "identifying variables" also arises in the context of the method discussed next.

6.2 Instrumental Variable Methods

Selection bias arises because of the correlation between the indicator variable for participation/non-participation in the program and the random error term in the outcome equation. The "instrumental variable" (IV) method to solving the selection bias problem, discussed by Heckman and Robb (1985) and Moffitt (1991) among others, centres on finding a variable (or variables) that influences selection into the program but does not influence the outcome of the program (and is thus not correlated with the random error term in the outcome equation). Because the instrumental variable is not correlated with the random error term, it can be used in the estimation without introducing bias. The formula for the IV estimator is given in Appendix C.

The search for IVs entails an in-depth investigation of the selection process. Personal characteristics of individuals would seldom suffice as instrumental variables because they are usually related to the outcome. For instance, level of education likely affects one's employability. Moffitt suggests that variation in the availability of treatment may yield a suitable variable. If a training program is available in one region but not another for reasons unrelated to the program's

intended outcome, region is a legitimate instrumental variable. This may be the case if the program is not available for political, bureaucratic or economic reasons.

In order to be a legitimate "instrument," the variable must be related to program participation/ non-participation but unrelated to the outcome(s) of the program. In some situations there may be numerous potential instrumental variables. In these circumstances, how should the analyst choose among these various potential IVs? The answer to this question is as follows. Each IV that is indeed unrelated to the outcome of the program (i.e. is uncorrelated with the random error term in the outcome equation) will yield unbiased estimates of the impact of the program. However, some IVs will yield more precise estimates of the impact of the program. Specifically, the more highly correlated is the IV with program participation/non-participation, the more precise will be the estimates of program impact. Thus the challenge in IV estimation is to find an instrumental variable that is highly correlated with program participation but uncorrelated with the outcome of the program. Unfortunately, it is often difficult to find variables that meet both these requirements, and therefore difficult to find good IVs among the many potential IVs.

As a way of further understanding the principles of IV estimation, consider the case of a program in which participation/non-participation is determined by random assignment. Suppose that each applicant is assigned as a participant P (non-participant NP) if the toss of a fair coin yields Heads (Tails). Then the indicator variable H (which equals unity if Heads and zero if Tails) is an ideal instrumental variable because H is perfectly correlated with the indicator variable for participation and because H is uncorrelated with the outcome of the program. Of course, in practice it is rare to have available such an ideal IV, but this example illustrates the characteristics one searches for when using this method.

This method thus has some features in common with the "identifying variable(s)" used in the two step method discussed above. In effect, both procedures require similar information. The main differences are: (i) the instrumental variable method is carried out in a single stage and does not therefore involve explicitly modelling the process of participation into the program; and (ii) the IV method produces estimates free of selection bias (if the model is properly specified) but does not provide an estimate of the magnitude of the selection bias, as is provided in the two step method[6].

Many examples of "instrumental variables" could be given. Moffitt (1991) gives the example of a government-funded health counseling program which, for reasons that are unrelated to the health needs of the populations in the two areas, funds the program in one area of a city but not in the other area. As a consequence, an indicator variable for the two areas of the city is unrelated to the health needs of the population in the two areas, but will influence the participation in the program. Another example, related to the assessment of public versus private schools discussed above, would be a measure of proximity to a private school for each of the students

---

[6] Also, the IV estimation procedure does not require the assumption of normally distributed random error terms in the first stage probit equation used in the two step procedure.

in the sample. Such proximity would be expected to influence the likelihood of attending a private school, but not the outcome of private schooling on student achievement. Appendix C has more on the IV method including equations.

6.3 Longitudinal Methods

The two step and instrumental variables methods discussed above can be implemented with post-program data alone (that is, cross-sectional data on participants and non-participants). However, if pre-program data on participants and non-participants are available, this should also be used and its incorporation in these methods will generally lead to more precise and more credible estimates of program impact. Longitudinal data follow the same individuals over two or more periods of time, and the longitudinal methods discussed here require at least one pre-program observation and at least one post-program observation on both program participants and non-participants.

The most common longitudinal estimator of program impact is the "fixed effects" or "difference-in-differences" estimator (sometimes also simply called the "differences" estimator). In the simplest case, in which there is one pre-program observation and one post-program observation, this approach proceeds as follows. First, take the difference between the post-program value of the outcome measure and the pre-program of the outcome measure for each participant and non-participant. This difference is thus a measure of how much change was observed in the outcome of interest between the period prior to the program and the period following the program. If the outcome of interest is earnings, as would be the case in assessing training programs, this would be the earnings gain or loss for each participant and non-participant. Next, take the difference between the average pre- versus post-program change for participants and the average pre- versus post-program change for non-participants (see Appendix C for equation). In the case of a training program, this is simply the difference between the average earnings gain or loss of participants and the average gain or loss of non-participants.

This simple estimator of program impact will be free of selection bias if selection into the program depends on unobserved person-specific "fixed effects," that is, factors that are specific (or unique) to each individual in the sample, but are constant ("fixed") over the time period of the analysis. For example, in the case of comparing coop and non-coop educational programs, such an unobserved factor could be how career-oriented the individual student is. If, as discussed previously, selection into coop and non-coop programs depends on how career-oriented the student is, and if this factor is not observed by the researcher (as will often be the case), the bias that would arise because of this unobserved factor will be removed by the use of the "difference-in-differences" estimator providing the extent of career-orientation of the student is constant over the sample period. Similarly, in the case of training programs, the "fixed effects" assumption is appropriate when unobserved person-specific factors such as ambition, labour force attachment and suitability for training are constant over time (but may vary across individuals). In many cases, such a "fixed effect" assumption will seem reasonable for such unobserved person-specific factors, although the validity of the assumption should be tested as discussed further below.

In summary, the difference-in-differences estimator will yield unbiased estimates of program impact — even in the presence of potential selection bias — when the source of potential bias is a correlation between participation/non-participation in the program and an unobserved factor which may differ across individuals but which is constant over time for each individual. If selection into the program takes this form, the simple difference-in-differences estimator is a straightforward way of dealing with selection bias.

More complicated longitudinal estimators are available for situations in which the assumption of constant or "fixed" person-specific effects is not appropriate. These generally require more than one pre-program and one post-program observation on each individual. For example, Moffitt (1991) discusses a "difference-in-differences in growth rates" estimator which is appropriate when the period-to-period change in the person-specific effect is constant over time. This estimator requires at least two pre-program and two post-program observations. Other types of longitudinal estimators are discussed in the context of training programs by Ashenfelter and Card (1985).

6.4 Specification Tests of Alternative Models

Three general classes of methods of dealing with selection bias have been outlined. Within each of these general classes, there are a number of variants on the basic procedure. The question which naturally arises is: Which of these methods should be employed, and under what circumstances?

In part, the answer to this important question is that the appropriate method should be determined by the researcher/evaluator, according to the circumstances of the program being assessed. A key aspect of being a good evaluator is being able to specify the model, including the methods for dealing with potential selection bias, appropriately. This aspect requires judgment, experience, and the ability to obtain information about the nature of the process by which participants are selected into the program.

Although factors such as judgment and experience are important, in most circumstances they are unlikely to be sufficient to enable the evaluator to know with reasonable certainty which method is most appropriate for dealing with selection bias. For this reason, it is important to employ a variety of specification tests which are available for determining which models or specifications are consistent with the data and which are not consistent. Examples of such specification tests are described in contributions by Heckman and Hotz (1989) and Moffitt (1991). Heckman and Hotz (1989) find, in the case of training programs, that use of a number of specification tests enables them to substantially narrow down the number of possible alternative forms which selection into the program may take, thus reducing substantially the range of possible estimates of program impact.

To date, such specification tests to assess the validity of alternative models have not been used as extensively as should have been the case. However, in part because of recent contributions to the evaluation literature, this use is increasingly becoming an important aspect of well-executed evaluations. An important side-effect of this trend is that evaluators are increasingly

required to devote thought and attention to the process by which selection into the program takes place (and therefore to the appropriate way to take account of potential selection bias), rather than to rely on a mechanical technique such as the Heckman two-step procedure.

6.5 Determinants of Program Participation

A final observation is that the more information that can be obtained about the process by which participants end up in the program and non-participants do not end up in the program, the more credible will be the estimates of program impact. As discussed previously, program participation depends on both observed and unobserved factors, and accounting for the influence of observed factors is much more straightforward and less uncertain than accounting for the role of unobserved factors. Thus acquiring richer data on the determinants of program participation is one of the most effective methods of dealing with potential selection bias.

In any evaluation there will always be some factors that are unobserved but which potentially could result in selection bias. In order to best take account of such possibilities, the greater the availability of rich qualitative information on the program and on participant and non-participant characteristics, the more able the evaluator is to choose the most appropriate method for addressing selection bias.

7.0 Selecting Outcome Measures

A critical step in a summative evaluation is to select the best measures for assessing outcomes. "An irrelevant or unreliable measure can completely undermine the worth of an impact assessment by producing misleading estimates" (Rossi and Freeman, 1993, p. 234).

Outcome measures relate to the impact the program is supposed to have, so suitable outcome measures focus on the program's objectives. In general, HRDC training programs have all or a subset of the following goals: higher educational achievement; improved transition to the labour market with enhanced employability and earnings; reduced dependency on passive income support; and improved work attitudes. These outcomes are easily quantified and would be treated as dependent variables in the econometric models (since it makes sense to assess the program in terms of its intended effects). Thus, good outcome (post-program) variables for training programs include education level, employment status (i.e., working or not), time spent working or in school, annual earnings, months spent on social assistance, weeks spent on EI, and attitudes towards work, education and passive assistance.

It is important to note that pre-program measures of the outcome are highly desirable for any quasi-experimental evaluation. Very occasionally, the program's management information system will have adequate pre-program data for participants and non-participants. For example, HRDC has complete and accurate data on EI use that are very useful for impact evaluations. HRDC's Service and Outcome Measurement System (SOMS) also has good longitudinal data on employment status, and time spent in school or work. Unfortunately, reliable pre-program data on social assistance use and attitudes are very rare. Nevertheless, this information should be gathered if possible to conduct an impact evaluation. Some pre-program information

can be gathered during a follow-up survey, although problems of accurate recall and lost records render the data imprecise. Other pre-program information, especially attitudes, is impossible to reconstruct after the program.

Collecting pre-program data on the sample through a "baseline survey" is the best recourse. The fundamental purpose of a baseline survey is to establish the pre-program characteristics of participants and non-participants in support of a future summative evaluation. A good baseline survey would feature questions that aim to establish what the person was doing in terms of the outcome variables (work, school, earnings, social assistance, EI, and so on) before the program. It should explore pre-program events each year for at least two years before the program began: for example, the number of months on social assistance during 1995, 1996 and 1997. A good section on attitudes, and a section on demographics (especially if the administrative system is unreliable or non-existent) should also be included in the questionnaire. Finally, the baseline survey should ask for at least two contacts (family or friends) who can help locate the individual for follow-up purposes, because the target group for training programs tends to be very mobile.

8.0 Selecting a Comparison Group

The most important aim in constructing comparison groups is to match as closely as possible the observed factors that might predispose a person to be successful in terms of the outcome measures. A clear example would be the influence of pre-program education on earnings: there is a well-established positive correlation between education and earnings levels (Lalonde, 1995). Ideally, then, one would want the treatment and comparison groups to be similar in terms of pre-program education level. Evaluators would thus want a comparison group with means and distributions as close as possible to the trainee sample in terms of education.

Heckman et al (1995) claim to have developed an effective matching methodology based on several variables (as reported in Heckman and Smith, 1995b). Central to creating truly comparable comparison groups, they assert, is information on "labour force status transitions," especially transitions from employment to unemployment and from outside the labour force into unemployment. Other key matching variables according to these scholars are age, education, marital status and family income.

Geographic location and time period (e.g., both groups were on social assistance during a similar period) are other important matching variables (see Friedlander and Robins, 1995). Which variables to use depends on what information is available and the purposes of the program and of the evaluation.

Several different methods are used to accomplish the matching. Eligibility matching, cell matching and statistical matching procedures are popular.

In eligibility matching, cases are selected from a representative sample of the population that meet the eligibility requirements for the training program. A relevant example would be HRDC's longitudinal file, which includes all EI recipients and the training courses they have taken. Both

trainees and non-trainees could be selected from this database (presuming the training focused on EI recipients).

In cell matching, also called stratified matching, individual observations in both samples are divided into cells defined by the traits that predict the outcome variable. For example, one might pre-assign sample members to cells based on region, date of EI application, age, and so on. Cells with no trainees are eliminated, those with few cases are combined. Some studies then weight variables to equalize distributions between groups (Fraker and Maynard, 1987).

To construct a statistical match (e.g., via the closest neighbour technique) comparison group, an individual is chosen from the eligible population for each trainee based on the closeness-of-fit on predicted scores on predicted outcome, or on closeness-of-fit on traits correlated with the outcome (Fraker and Maynard, 1987). Background characteristics of each trainee are matched against those of each non-trainee; the non-trainee most closely resembling the trainee is selected.

Fraker and Maynard (1987) concluded that the general population is an inadequate source of comparison samples. Friedlander and Robins (1995) reported that statistical matching produced mediocre improvements in the accuracy of quasi-experimental estimates. Dickinson, Johnson, and West (1987) reported that the type of matching procedure (cell matching or statistical matching) did not affect results. Riddell (1991), in his review of training evaluations, agreed that the specific matching procedure used did not appear to affect impact estimates, although ". . . data bases which focus on members of the target groups may increase the accuracy of quasi-experimental estimates of programme impact."

Therefore, considering that individual characteristics can be controlled for in the model anyway, it seems safe to conclude that *the choice of matching procedure is not as important as ensuring the comparison group would meet the eligibility requirements for the program.*

9.0 Applying the Lessons: Choosing Comparison Groups for EBSM

Lessons from previous quasi-experimental evaluations can assist the process of drawing regional EBSM comparison groups:

- To date, there is little evidence that the matching procedure used makes much difference to the accuracy of estimates from quasi-experiments. Thus, the simplest course would be to ensure the comparison group meets the eligibility requirements for the training program. *At the very least the comparison group should be drawn from the same population as the treatment group.* EBSM programs, which are primarily aimed at EI recipients, could use the HRDC longitudinal file for choice of comparison group subjects, for example. The NESS database may also be a useful source of comparison group members.

- Individual components of EBSM have specific eligibility criteria that should be taken into account when selecting a comparison group. For instance, targeted wage subsidies are largely focused on those who face particular obstacles to employment such as disabilities. The sample frame for the comparison group should come from the subset of the EI population who share these obstacles.

- Moving one step beyond simple eligibility matching is advisable in the case of regional EBSM programs: it seems clear that comparison groups should at least come from the same region. In the case of Transitional Jobs Fund, the comparison group should be chosen from the same communities as the treatment group (i.e., those with unemployment rates of 12% or higher). If desired, and assuming the requisite information is available from the database used, other potentially important matching variables are labour force status changes, age, education, marital status, family income, and time period.

- Sampling strategy depends on the matching procedure employed. The easiest approach — and one that would be no less satisfactory than more complex ones judging by the literature — would be to first limit the population to those in the region who were on EI during the period that the program was in operation (plus any other eligibility factors unique to each EBSM component). Then a simple random sample could be selected. A more precise matching strategy would be more difficult to implement, but by no means impossible. For example, if one wanted to match by sex, age and education, a series of dummy variables (0-1) could be set up: men=0, women =1; under 30 years old=0, 30+=1; less than high school=0, high school gradute=1. A three-digit variable could then be computed for each participant and non-participant with the first digit representing sex, the second age, the third education (e.g., a 29 year old female high school graduate = 101). The software program could then be used to select at random, a match (or more than one match) for each participant.

- As for sample size, standard formulas should be employed, keeping in mind that they calculate margins of error for final sample size rather than the original sample size. Appendix D summarizes how to choose an appropriate sample size.

10.0 Conclusion

This paper has presented the basics of quasi-experimental evaluation design, with a focus on considerations for controlling selection bias and for choosing comparison groups. Quasi-experimental designs are usually the best approach to use for evaluating social programs because one-group designs are inadequate for establishing cause and effect, and two-group experimental designs are very often impractical. Complex analytical procedures are required to isolate the effect of the program from many other potential causes under this design, however.

Appendix A: True Experiments Versus Quasi-Experiments

There has been a lively debate for over a decade on whether true experiments or quasi-experiments represent the superior design. This appendix will present each side's argument.

The Experimental Camp

Those favouring the experimental approach question the validity of highly sophisticated non-experimental approaches for assessing the impact of training and employment programs because of the apparent difficulty in obtaining reliable estimates of the impact of labour market programs. They claim that economists using quasi-experimental techniques have had little success in isolating program effects (i.e., removing the "selection bias").

Scholars such as Ashenfelter and Card (1985), Barnow (1987), and LaLonde and Maynard (1987) contend that results of dozens of econometric studies were so varied that there may be no sound way to adequately measure program effects short of an experimental evaluation with random assignment to training or control groups.

The findings of Lalonde (1986) and Lalonde and Maynard (1987) were exceptionally damaging to the case for non-experimental designs. To assess the accuracy of such designs, they compared the results from a true experiment of the National Supported Work (NSW) Demonstration Project to those derived from various widely used non-experimental procedures to see if they could accurately estimate the true program impacts. They concluded that "there does not appear to be any formula [using non-experimental methods] that researchers can confidently use to replicate experimental results of the Supported Work Program." Using the same evidence, Fraker and Maynard (1987) concluded:

> This analysis demonstrated that results may be severely biased depending on the target population, the comparison group selected, and/or the analytic model used. More importantly, there is at present no way to determine a priori whether comparison group results will yield valid indicators of the program impacts (p. 216).

More recent work confirms the chief conclusions drawn from this work – that quasi-experimental estimators are biased and sensitive to minor changes in model specification. Friedlander and Robins (1995) assessed two conventional quasi-experimental strategies against experimental data from four social assistance reform experiments: comparing the treatment group in one locale to a comparison group in another locale; and comparing outcomes of the treatment group with a pre-program comparison group in the same area. They concluded that the non-experimental estimates were "usually quite different from the experimental estimates," especially for the comparison samples drawn from different areas. They also studied two statistical

techniques for improving the accuracy of quasi-experimental estimates: statistical matching to produce closely matched comparison groups; and "specification tests," to statistically assess the econometric model employed to determine if the estimates it yields are accurate[7]. Neither strategy markedly improved the accuracy of the non-experimental estimates.

Further support for the experimental method comes from Greenberg and Wiseman (1992), who stated that fifteen years of experimentation (e.g., income maintenance experiments, Supported Work Demonstration) have demonstrated that random assignment was "a methodologically superior approach to program evaluation," and provided evidence that such studies were feasible. Moreover, the work on the Omnibus Budget Reconciliation Act evaluations (by the Manpower Demonstration Research Corporation) convinced decision-makers at the U.S. Department of Health and Human Services — which usually funds evaluations of training programs — that random assignment should be used as the foundation of social assistance reform evaluations. MDRC's evaluations were generally considered excellent, helping to establish random assignment as the procedure of choice.

Another advantage of experimentation is that the results are understandable and convincing to policy makers (Burtless, 1995). Without the complicated qualifications associated with quasi-experiments, analysts can present straightforward experimental findings such as "the program raised annual earnings of participants by $1,000." This simplicity makes it more likely that policy makers will use the evaluation findings. "They do not become entangled in a protracted and often inconclusive scientific debate about whether the findings of a particular study are statistically valid. Politicians are more likely to act on results they find convincing" (Burtless, 1995, p.67).

Finally, a National Academy of Sciences panel recommended the following conditions as necessary (but not sufficient) for quality research: *the use of random assignment to group*; reasonable operational stability of the program prior to final assessment; adequate sample coverage and low rates of sample attrition; outcome measures that well represent program objectives, both immediate and longer-term; and a follow-up period that allows time for program effects to emerge or decay (Gueron and Pauly, 1991).

These studies, among others, have convinced many evaluators that experimental estimators are better than quasi-experimental ones (e.g., Burtless 1995; and Friedlander and Robins 1995).

---

[7] For example, one can see if the model correctly predicts no differences in outcomes between trainees and non-trainees during the period before the program began. If the model finds statistically significant differences between groups before the program, it should be dropped since it failed the specification test.

The Quasi-Experimental Camp

Others, however, have contested the view that experimental designs are superior to non-experimental designs. In the forefront of this group are James Heckman and his colleagues (e.g., Heckman, Hotz, & Dabos, 1987; Heckman and Smith, 1995). They argue that with a sufficiently rich data set and appropriate econometric modeling techniques, it is possible to arrive at reliable estimates of impact.

To mount a serious challenge to the emerging consensus that experimentation was the better route, they had to surmount the conclusion of Lalonde and his associates that no available quasi-experimental method produced estimates close to those of the unbiased experimental estimates. Heckman and Hotz (1989) applied some simple specification tests to models estimated on data from the NSW experiment and found that the most inaccurate models could be rejected, leaving a subset of models that produced impact estimates close to those of the experimental results. Heckman and Smith (1995) presented some cogent arguments that undermine to some extent Lalonde's findings: sample sizes were too small and insufficient geographical data were available to place comparison group members in the same local labour market as participants; there was only one year of pre-program data available, ruling out potentially effective econometric strategies (and leaving the estimates subject to the "Ashenfelter Dip"[8]); the studies did not use a variety of standard specification tests; and there have been important advances in non-experimental methods since the these studies were done. For instance, Heckman and Smith (1995b) found that labour force dynamics (i.e., movements between employment, unemployment and into and out of the labour force), as well as other contributing factors such as age, education, marital status and family income, can be used to form comparison groups that are "virtually identical" to treatment groups, thereby controlling for selection bias.

Heckman and Smith (1995) also raise some serious objections to the experimental method, both theoretical and empirical:

- Randomization may alter the pool of persons eligible for the program or change the behaviour of participants, the so-called "randomization bias." For instance, in order to form a control group, occasionally the pool of potential participants must be expanded, usually by relaxing some eligibility criteria.

- If close substitutes for the experimental treatment are available, the assumption that the control group receives no treatment is annulled to the extent members take advantage of the training, generating a "substitution bias." For instance, some training programs consist of buying seats at community colleges; non-participants can choose to attend the same college course if they pay on their own or find another source of funding.

---

[8] This refers to the phenomenon that the earnings of training program participants tend to dip just before they enter training, because unemployment is often the impetus to take the course; thus, before and after difference estimators will tend to overestimate the effect of the program

- Experimental data cannot answer many of the questions of central interest to policy makers, including the median impact of the program and the proportion of participants with a positive (or negative) impact from the program. No parameters that depend on the joint distribution of outcomes in the treatment and control groups can be estimated. "Only if the evaluation problem is defined exclusively in terms of means can it be said that experiments provide a precise answer" (Heckman and Smith, 1995, p. 22).

- Experimental evaluations tell policy makers whether the programs work or not. They do not typically shed any light on why a program worked or did not work.

- Institutional factors make it hard to carry out random assignment at all in some cases and difficult to do an optimal job in others. Staff can subvert the process because the are opposed to random assignment. Costs can militate against making the selection at the optimal point in the decision process[9]. Multi-stage randomization, which is expensive and disruptive to the program, may be required to generate estimates of the impacts of differ-ent services.

- Complementary non-experimental analyses are often required to overcome the short-comings of experimental models.

Burtless (1995) agrees that experimental evaluations have such shortcomings, but counters that quasi-experiments also share some of them (e.g., quasi-experiments can also be costly and disruptive); plus they are "usually plagued by more serious statistical problems than those that occur in randomized trials." Moreover, he adds, that nothing inherent to the experimental design precludes researchers from using non-experimental methods to analyze data.

Conclusion

The jury is still out as to whether quasi-experimental designs can adequately control selection bias. It is safe to conclude that experimental designs are superior in this critical respect. But the many problems associated with experiments render them impractical for many if not most evaluations. Quasi-experimental designs are often the best practical approach to take for evaluations of training programs.

[9] Randomization can occur at any step of the process: participant becomes eligible, becomes aware of the program and his/her eligibility for it, applies, is accepted, is assessed by staff, is assigned to particular services, begins receiving the services, and completes the program. The optimal place for randomization depends on the evaluation questions at issue. For determining mean impact, the assignment should take place as close as possible to the commencement of training to minimize attrition.

Appendix B: Brief Mathematical Representation of Quasi-Experiments

The problem of trying to evaluate the impact of a social program in a non-experimental setting may be represented as follows (Moffitt, 1991):

$$Y^{**}_{it} = Y^{*}_{it} + \alpha$$

$$\alpha = Y^{**}_{it} - Y^{*}_{it}$$

where

Y^{*}_{it} = level of outcome variable for person i at time t if he had not participated

Y^{**}_{it} = level of outcome variable for same person at same time if he participated previously

Evaluations aim to estimate α, the treatment effect. That is, we wish to estimate for those who have participated what Y would have been had they not participated. Clearly, we cannot know Y^{*}_{it} since these individuals ***have*** used the program. So we substitute Y^{*}_{it} of non-participants to estimate α:

$$\alpha = E(Y^{**}_{it} | d_i = 1) - E(Y^{*}_{it} | d_i = 0)$$

where $d_i = 1$ if person i has participated

$d_i = 0$ if person i has not participated

and | denotes "conditional on" so that the first term on the right hand side of the above equation is the mean value of Y for participants and the second term on the right hand side is the mean value of Y for non-participants.

In words, we estimate the treatment effect by estimating the expected value (E) of Y, say annual earnings, for those who have participated in a training program, and subtracting the expected value of Y for those who have not. Only if pre-program $E(Y^{**}_{it})$ for participants equals pre-program $E(Y^{*}_{it})$ for non-participants will there be no bias. But this will seldom be the case because of selection bias.

Appendix C:
Dealing With Selection Bias

Undoubtedly the most popular approach to dealing with selection bias is the Heckman (1979) *two-stage approach*. The first stage involves modeling selection into the program. Usually this takes the form of a single equation explaining program participation/non-participation[10]:

$$P = \beta X + U$$

where P is a dummy variable (1 for participants and 0 for non-participants), X is a set of all observed factors that may account for participation in the program (e.g., age, sex), and U is a random error term which is assumed to be normally distributed to take account of unobserved factors that influence participation in the program. From this equation, the inverse of Mill's ratio is computed, which is then inserted into a second stage outcome equation to estimate program impact (usually via ordinary least squares):

$$Y = \beta X + \alpha P + \delta M + U$$

where Y is the outcome of interest. X is a vector of observed variables, P is the participation dummy, and M is the inverse of Mill's Ratio. If the assumptions underlying the model are correct, the Heckman procedure removes the selection bias (δ), thereby producing an unbiased estimate of program impact (α). (The measure of program impact is the estimated coefficient on the indicator variable for participation/non-participation in the program.) If this equation were estimated by ordinary least squares without the inclusion of the selection bias correction term, the estimates would potentially be biased. However, if the model is properly specified, the addition of the "selection bias correction" variable removes this potential bias, thus giving unbiased estimates of program impact.

Another powerful means of controlling for differences between groups is called the *differences-in-differences* method. Longitudinal data are collected for key outcome measures 3/4 e.g., earnings, social assistance use. To account for the differences in the participant and non-participant samples a longitudinal estimator of program impact is employed; such estimators take account of the level of the outcome variable prior to and after the program, in contrast to cross-sectional estimators which use data on post-program outcomes alone. This estimator uses the pre- vs. post-program change in the outcome variable for non-participants as an

[10] In most cases, the equation is estimated as a "probit model" which is appropriate if the random term in this equation is normally distributed. (A probit is a measurement of probability based on deviations from the mean of a normal frequency distribution. It is analogous to multiple regression but with a dischotomous dependent variable.) However, two step prodedures for situations in which the assumption of normally distributed random terms is unlikely to hold are available (through not yet widely used).

estimate of the change that would have occurred for participants in the absence of the program. The estimated average program impact is then the difference between the pre- vs. post-program change in the outcome variable for participants and the pre- vs. post-program change in the outcome variable for non-participants. This permits a determination of the incremental impact of the program by controlling for biases caused by unobserved individual differences. A multivariate analysis can then show how the size of the differences-in-differences estimate of program impact varies according to various individual and program characteristics.

In equation form (Moffitt, 1991):

$$Y = E(Y^{**}_{it} - Y^{*}_{i,t-1} \mid d_i = 1) - E(Y^{*}_{it} - Y^{*}_{i,t-1} \mid d_i = 0)$$

where t = the posttreatment point, t-1 = pretreatment point, and

$Y^{*}_{it} - Y^{*}_{i,t-1}$ = change in Y^{*}_{it} from t-1 to t if treatment not received

$Y^{**}_{it} - Y^{*}_{i,t-1}$ = change in Y^{*}_{it} from t-1 to t if treatment received

Instrumental variable (IV) methods are widely used in situations in which ordinary least squares (OLS) estimates may be biased due to a correlation between one or more of the explanatory variables and the random error term in the model. In the evaluation/selection bias context, such potential bias arises because of the possible correlation between the participation/non-participation variable and the random error term in the outcome equation. This potential bias can be removed if one or more "instrumental variables" are available and included in the model.

The basic model of program outcome or impact is:

$$Y = \beta X + \alpha P + U \quad (1)$$

This can be written as:

$$Y = CW + U \quad (2)$$

where $C = (\beta\ \alpha)'$ and $W = (X\ P)$ using matrix notation.

The least squares estimator of (2) is given by:

$c = (W'W)^{-1} W'Y$ where $(W'W)^{-1}$ is the inverse of the matrix (W'W). In general this estimator is biased because of the correlation between W and U (i.e. E{W'U} does not equal zero, where E{ } represents the expectations operator).

The IV estimator of (2) is given by:

$c^* = (Z'W)^{-1} Z'Y$ where Z is the matrix of instrumental variables. This estimator is in general unbiased because Z and U are uncorrelated, i.e. E{Z'U} equals zero if Z is an appropriate instrument.

Appendix D: Determining Appropriate Sample Size

What sample size should be chosen for a particular survey? It depends mainly on tolerable error, population size, the importance of particular subgroups, anticipated level of non-response, and how much money is available.

"Tolerable error" refers to the margin of error for the survey. Whenever the results of polls are reported in the news, the margin of error – e.g., plus or minus 3%, 19 times in 20 – is included. The margin of error tells the reader how accurate the poll's findings are. It is based on the "standard error," the measure of how much the sample mean differs from the population mean. The margin of error adjusts the standard error to account for any potential differences between the sample and the population via calculation of a "confidence interval" for the population mean. Traditionally, a 95% confidence interval is used (i.e., 19 times in 20). The tolerable margin of error is usually between 3% and 5% (much lower and the costs of the survey begin to rise dramatically).

The traditional formula for large population sizes is $n = 1.96^2\ p(1-p)/SE^2$, where n is sample size to be calculated, SE is the tolerable standard error, and p is the proportion having the characteristic being measured and (1-p) is the proportion who lack it (e.g., if 48% said yes, 52% must have said no). The 1.96 figure reflects the choice of a 95% confidence interval (in a normal distribution, 95% of the area under the curve is within 1.96 standard deviations of the mean). For example[11], if a margin of error of ± 3%, 19 times in 20 was tolerable, the following sample size would be required:

$$n = 1.96^2\ (.5*.5)/.03^2 = 1{,}068$$

Population size is a consideration only when it falls below 100,000 or so. Below that, something called the "finite population correction factor" must be used to determine sample size. The correction factor = $(N - n\ /N - /1)^{1/2}$, where N is the population size and n is the sample size.

Algebraically entering this factor into the sample size equation, yields:

$$n = (1.96^2\ p(1-p)N)/\ (1.96^2\ p(1-p)) + (N-1)SE^2$$

[11] By convention, p and 1-p are set to the most conservative level – .5 for each.

For example, if an evaluator wanted to learn how many EI clients to survey from a sample frameof 2,500, with an error rate of ± 3% at 95% level of confidence:

$$n = (1.96^2 * .25(2500))/ (1.96^2 * .25) + 2499(.0009) = 749$$

The sampling error associated with subgroups will be higher than that for the whole sample, because there are obviously fewer cases. A rule of thumb is that there should be a minimum of 100 individuals in any major subgroup that will be analyzed separately. This will achieve at least a ± 10% margin of error for each major stratum, the maximum tolerable (Rea and Parker, 1992).

Note that when choosing a sample size, there will always be some people in the sample who can't be located, or who will refuse to cooperate. Allowances must be made for anticipated non-response. For a *final* sample size of 1,000, with a 50% response rate, the *initial* sample size must be 2,000. Given a fixed budget, there is always a tradeoff between the initial sample size and the effort to reduce non-response. Too often a large initial sample is chosen and too little effort is expended in reducing non-response, with consequent effects on total error.

References

Ashenfelter, O. and D. Card (1985) Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics.* 67:648-660.

Barnow, B. (1987) The impact of CETA programs on earnings: a review of the literature. *Journal of Human Resources.* 22:157-93.

Burtless, G. (1995) The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives.* 9(2): 63-84.

Campbell, D.T. & J.C. Stanley (1971) *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally & Co.

Dickinson, K.P., Johnson, T.R., and R.W. West (1987) An analysis of the sensitivity of quasi-experimental net impact estimates of CETA programs. *Evaluation Review.* 11:452- 472.

Fraker, T. and R. Maynard (1987) The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources.* 22:194-227.

Friedlander, D. & P.K. Robins (1995) Evaluating program evaluations: new evidence on commonly used nonexperimental methods. *American Economic Review.* 85(4):923-937.

Greenberg, D. & M. Wiseman (1992) What did the OBRA demonstrations do? In C.F. Manski & I. Garfinkel (Eds.) *Evaluating Welfare and Training Programs.* Cambridge: Harvard University Press.

Gueron, J.M. & E. Pauly (1991) *From Welfare to Work.* New York: Sage. Heckman, J.J. (1979) Sample selection bias as specification error. *Econometrica* 47: 153- 161.

Heckman, J.J., V.J. Hotz, and M. Dabos (1987) Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review.* 11:395-427.

Heckman, J.J.& V.J. Hotz (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association.* 84:862-877.

Heckman, J. and R. Robb (1985) Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, edited by J. Heckman and B. Singer, 156- 246. Cambridge: Cambridge University Press.

Heckman, J.J & J.A. Smith (1995) Assessing the case for social experiments. *Journal of Economic Perspectives*. 9(2):85-110.

Heckman, J.J & J.A. Smith (1995b) Ashenfelter's dip and the determinants of participation in a social program: implications for simple program evaluation strategies. Unpublished manuscript, University of Chicago.

Heckman, J.J., H. Ichimura, J.A. Smith & P. Todd (1995) Nonparametric estimation of selection bias using experimental data. Unpublished manuscript, University of Chicago.

Lalonde, R.J. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*. 76(4):604-620.

Lalonde, R.J. (1995) The promise of public sector-sponsored training programs. *Journal of Economic Perspectives*. 9(2):149-168.

Lalonde, R.J. and R. Maynard (1987) How precise are evaluations of employment and training programs: Evidence from a field experiment. *Evaluation Review*, II:428-451.

Mark, M.M. & T.D. Cook (1984) Design of randomized experiments and quasi-experiments. In L. Ruttman (ed.) *Evaluation Research Methods*. Beverly Hills, CA.: Sage.

Moffit, R. (1991) Program evaluation with nonexperimental data. *Evaluation Review*, 15:291-314.

Riddell, C. (1991). Evaluation of manpower and training programmes: The North American experience. In OECD (Ed.) *Evaluating Labour Market and Social Programmes*. Paris:OECD.

Rubin, D.B. (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*. 74:318- 328.

Patton, M.Q. (1982) *Practical Evaluation*. Beverly Hills, CA.: Sage. Posavac, E.J. & R.G. Carey (1980) *Program Evaluation: Methods and Case Studies*. Englewood Cliffs, N.J.: Prentice-Hall Inc.

Rea, L.M. & R.A. Parker (1992) *Designing and Conducting Survey Research*. San Francisco: Jossey-Bass.

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation: A Systematic Approach* (5th ed.). Newbury Park, California: Sage.

Stufflebeam, D.L. & A.J. Shinkfield (1985) *Systematic Evaluation.* Boston: Kluwer Academic Publishers.

Évaluation quasi-expérimentale

Évaluation et développement des données
Politique stratégique
Développement des ressources humaines Canada

Janvier 1998

SP-AH053F-01-98
also available in English

Remerciements

Les Services d'évaluation de DRHC utilisent normalement une méthode de recherche quasi expérimentale (méthode des groupes témoins) pour estimer l'incidence d'une intervention. Afin de s'assurer que la méthode des groupes témoins convienne aux évaluations ultérieures des programmes et services offerts dans le cadre de la partie II de la Loi sur l'assurance-emploi, l'étude ci-jointe, intitulée « Évaluation quasi-expérimenatale », a été menée.

Nous voudrions remercier le Dr Robert Power, de Power Analysis Inc., à London, en Ontario, et le Dr W. Craig Riddell, de la faculté d'économie, Université de la Columbie-Britannique, d'avoir étudié et analysé les arguments pour et contre la méthode des groupes témoins et d'avoir préparé le rapport des constatations à partir d'une analyse documentaire approfondie.

Série

Développement des ressources humaines Canada (DRHC) a une longue tradition en matière d'évaluation de programmes. Au fil des ans, la Direction générale de l'évaluation et du développement des données (EDD) de DRHC a produit de nombreuses évaluations sur un large éventail de programmes ministériels. Elle a ainsi acquis un savoir-faire et une expérience qui lui ont permis d'améliorer ses techniques et méthodes d'évaluation.

Récemment, il a été nécessaire de revoir la façon dont sont réalisées les évaluations afin de tenir compte des changements dans la mise en oeuvre de certains des programmes de DRHC. Les bureaux régionaux de DRHC participent maintenant plus directement et activement à l'évaluation des programmes.

EDD est donc fière de lancer une nouvelle série de publications intitulée « Les outils de l'évaluation » destinée à accroître les compétences et les connaissances en matière d'évaluation dans l'ensemble de DRHC et de ses organismes partenaires en offrant de l'information pertinente concernant la conception, la planification et la réalisation d'une évaluation. La série permet de regrouper l'information disponible sur un sujet donné dans de courts rapports axés sur l'évaluation des programmes de développement des ressources humaines.

Le domaine de l'évaluation suscite de plus en plus d'intérêt et évolue constamment. N'hésitez pas à nous faire part de vos commentaires sur la série ou sur une étude en particulier.

La première étude de cette nouvelle série met l'accent sur l'approche quasi-expérimentale en évaluation.

Table des matières

1.0 Définition de l'évaluation **3**
2.0 Les deux principaux types d'evaluation **3**
3.0 Les modèles d'évaluation sommative **4**
4.0 Types de formules **6**
5.0 Exemples de biais de sélection **10**
6.0 Méthodes de correction du biais de sélection **12**
6.1 Méthodes de correction en deux étapes 13
6.2 Méthodes des variables instrumentales 15
6.3 Méthodes longitudinales 17
6.4 Tests de caractérisation d'autres modèles 18
6.5 Facteurs déterminants de la participation au programme 19
7.0 Choix des mesures des résultats **19**
8.0 Choix d'un groupe témoin **20**
9.0 Application des enseignements : le choix des groupes témoins pour le PEMS **22**
10.0 Conclusion **23**

Annexe A : Expérience véritables et quasi-expériences **25**
Annexe B : Brève représentation mathématique des quasi-expériences **31**
Annexe C : Comment tenir compte de l'effet du biais de sélection **33**
Annexe D : Détermination de la juste taille de l'échantillon **37**

Bibliographie **39**

Évaluation quasi-expérimentale

Gestionnaire de programme : Pourquoi devons-nous évaluer notre programme ? Nous savons très bien ce qui se passe dans notre programme et avec nos clients et nous savons que nous avons beaucoup de succès.

Évaluateur : Parce que vous ne savez jamais si c'est votre programme ou un autre facteur qui vous a fait obtenir le succès que vous revendiquez.

Gestionnaire de programme : Mais c'est sûrement notre programme. Pourquoi pensez-vous qu'autant de gens trouvent un emploi aussi rapidement ?

Évaluateur : C'est peut-être en raison de l'expansion économique rapide que nous connaissons présentement.

Gestionnaire de programme : Il n'en est rien. De toute façon, nous savons que nous devons faire évaluer notre programme. Mais pourquoi devons-nous nous évertuer à trouver un groupe témoin pour faire une évaluation ?

Évaluateur : Disons que, six mois après la fin de votre programme de formation, 70 % des personnes qui ont reçu une formation travaillent. Est-ce là une preuve du bon fonctionnement de votre programme ?

Gestionnaire de programme : Je dirais que oui. Nous aimerions avoir un résultat supérieur à 70 % - nous croyons même que nous faisons encore mieux que cela - mais je dirais que, si 70 % des personnes qui ont reçu une formation travaillent, notre programme paraîtra très bien, étant donné surtout les obstacles auxquels se heurtent beaucoup de nos clients lorsqu'ils viennent nous voir.

Évaluateur : Quelle proportion de ces personnes travailleraient maintenant si elles n'avaient pas suivi votre programme de formation ?

Gestionnaire de programme : Je ne sais pas exactement, mais ce ne serait pas une proportion aussi élevée que 70 %, je puis vous l'affirmer.

Évaluateur : Mais vous n'êtes vraiment pas sûr de cela, car tout aussi bien 80 % pourraient travailler maintenant si elles n'avaient pas suivi la formation.

Gestionnaire de programme : Impossible. Vous ne savez pas à combien d'obstacles nos clients se heurtent lorsqu'ils viennent nous voir. Nous offrons un service précieux et nous aidons vraiment nos clients.

Évaluateur : C'est peut-être le cas, mais vous n'avez rien démontré. Or les promoteurs du programme veulent savoir avec certitude quel a été votre degré de réussite. Ils veulent savoir s'ils en ont vraiment pour leur argent.

Gestionnaire de programme : C'est le cas, je peux vous l'assurer.

Évaluateur : Disons donc que vous faites de bonnes choses, que les personnes qui bénéficient de votre programme sont en effet plus aptes à trouver un emploi que s'ils n'avaient pas reçu une formation. Quelles sont les répercussions de votre programme ? La moitié n'auraient-elles pas trouvé un emploi de toute façon ? Le tiers ? Les deux tiers ? Vous ne pouvez répondre à cette question à moins de faire une évaluation qui comprend un groupe comparable de personnes qui n'ont pas suivi votre programme de formation.

Gestionnaire de programme : Même si la moitié d'entre elles avaient trouvé un emploi sans la formation, ne vaut-il pas la peine d'amener cette proportion à 70 % ?

Évaluateur : Je ne sais pas. Combien en a-t-il coûté pour ce 20 % supplémentaire ? Quelle sera la durée des effets de la formation ?

Gestionnaire d e programme : Notre programme en vaut le coût...

Cet entretien fictif illustre bien les perceptions différentes des gestionnaires et des évaluateurs de programmes lorsqu'il s'agit de juger de la valeur d'un programme. Le programme, c'est la vie et c'est le travail du gestionnaire à tous les jours : il y tient et il travaille fort pour en assurer la réussite; il croit fermement qu'il fait un bon travail; et il est à juste titre indigné s'il entend dire que tel n'est pas le cas.

L'évaluateur n'a habituellement aucun lien avec le programme et surtout aucun intérêt dans sa survie (ce qui l'amène parfois à sous-estimer la menace qu'une évaluation peut faire planer sur les gestionnaires et les employés d'un programme). Il sait que le gestionnaire a beaucoup investi dans son programme et que sa propre évaluation du programme - même s'il dispose de données fiables - n'est pas reconnue par les promoteurs du programme comme une indication valable du fait que le programme atteint ses objectifs et en vaut le coût. L'évaluateur sait aussi que de nombreux facteurs différents non liés à la conception du programme peuvent avoir une incidence sur les résultats de tout programme social et peuvent facilement susciter des conclusions injustifiées.

Le présent document est fondé sur le postulat que <u>seule</u> une bonne évaluation peut établir des preuves convaincantes de l'efficacité d'un programme dans l'atteinte de ses objectifs et que <u>seule</u> une évaluation qui comporte un groupe témoin peut procurer, sur l'effet d'un programme, des données estimatives non contaminées par l'influence d'autres facteurs qui peuvent aussi avoir un effet sur les résultats. Le document résume les éléments fondamentaux de l'évaluation, en mettant en relief ce qui est habituellement la meilleure méthode pratique, soit le « modèle quasi expérimentale ». Il est rédigé en langage non technique pour les gestionnaires qui n'ont pas ou qui ont peu de connaissances du domaine de l'évaluation, mais il comprend aussi des

analyses plus approfondies dans des annexes pour ceux qui s'intéressent aux aspects plus techniques.

Le document commence par une brève introduction à l'évaluation, c'est-à-dire une large définition et un résumé des deux principaux genres d'évaluation. Suit un exposé sur le modèle de l'évaluation, en commençant par les raisons pour lesquelles l'évaluateur doit se soucier de retenir le bon modèle. Sont ensuite présentés les modèles les plus courants d'évaluation, ainsi que les avantages et les inconvénients de chacun. Nous expliquons pourquoi les évaluations qui ne comportent pas de groupes témoins sont loin d'être les meilleures puisqu'elles font abstraction des causes extérieures. Comme il faut dans toute bonne évaluation essentiellement comparer ce qui est arrivé aux clients du programme et ce qui serait arrivé s'ils n'avaient pas participé au programme, l'évaluation faisant appel à un seul groupe ne peut à toutes fins utiles donner lieu à une évaluation sommative sérieuse.

Nous donnons ensuite des précisions sur le modèle de l'évaluation quasi expérimentale : ses fondements théoriques et des considérations pratiques. La notion de biais de sélection est expliquée et les modèles économétriques visant à tenir compte de son effet sont présentées. Nous parlons aussi de la fiabilité des données économétriques.

Il est ensuite question du mode de sélection des groupes témoins. Nous résumons les diverses techniques d'appariement tout en évaluant leurs points forts et leurs points faibles relatifs. Nous examinons du même coup les variables possibles à utiliser dans l'établissement de l'appariement. Enfin, après avoir bien expliqué en général l'évaluation quasi expérimentale, nous commençons à appliquer aux tâches plus précises les enseignements tirés qui consistent à déterminer les meilleures techniques d'échantillonnage aux fins de l'établissement des groupes témoins régionaux pour les programmes des prestations d'emploi et des mesures de soutien (PEMS), ainsi que les meilleures variables à utiliser dans l'échantillonnage.

1.0 Définition de l'évaluation

Il existe de nombreuses définitions différentes de l'évaluation. Voici l'une des meilleures, parce qu'elle en relève les aspects les plus importants :

> « L'évaluation est un ensemble de méthodes, de techniques et de qualités de perspicacité permettant de déterminer si un service offert à des personnes est nécessaire et susceptible d'être utilisé, s'il est dispensé de la façon prévue et s'il est effectivement utile aux personnes. » (Posavac et Carey, 1980, p. 6).

Cette définition englobe les deux principaux types d'évaluation : l'évaluation des processus et l'évaluation sommative.

2.0 Les deux principaux types d'évaluation

Même si la littérature décrit plus d'une centaine de divers genres d'évaluation (voir Patton, 1982), la très grande majorité se résume à deux types : celles qui visent à déterminer si le programme a été mis en œuvre selon les intentions, et celles qui servent à établir la mesure

dans laquelle les objectifs ont été atteints (c.-à-d. ses effets). On appelle le plus souvent le premier type « évaluation des processus », mais parfois aussi « évaluation formative ». Le deuxième type s'appelle « évaluation sommative », mais aussi évaluation de l'effet, du résultat ou de l'efficacité.

Évaluation des processus — *Comment fonctionne le programme et comment l'améliorer ?* Dans l'évaluation des processus, trois questions fondamentales sont posées : (1) dans quelle mesure un programme atteint-il la population visée; (2) la prestation du service correspond-elle ou non à la conception du programme; (3) quelles sont les ressources utilisées. (Rossi et Freeman, 1993)

Il s'agit principalement d'indiquer aux gestionnaires si le programme est réalisé conformément aux intentions et d'une manière efficiente. L'évaluation est habituellement assortie de conseils visant la modification du programme pour qu'il atteigne ses objectifs. Avec ces renseignements, on peut modifier le programme pour qu'il soit exécuté selon les intentions, ou on peut modifier le plan comme tel si on constate qu'il présente des lacunes.

Évaluation sommative — *Le programme atteint-il ses objectifs ?* Dans l'évaluation sommative, on cherche à déterminer l'effet du programme, c.-à-d. la mesure dans laquelle il atteint ses objectifs et répond aux besoins du groupe cible. L'évaluation doit aussi comporter des conseils visant la modification du programme pour qu'il réponde mieux aux besoins de sa clientèle et que son rapport coût-efficacité soit amélioré. (Stufflebeam et Shinkfield, 1985)

Dans le reste de ce rapport, nous nous concentrerons sur l'aspect sommatif, en particulier l'évaluation sommative selon les modèles quasi expérimentales. Avant toutefois d'aborder les complexités du modèle quasi expérimentale, nous établissons le contexte en exposant les divers types d'évaluations sommatives.

3.0 Les modèles d'évaluation sommatif

Il n'existe pas un seul bon modèle pour l'évaluation de l'effet. Il s'agit de trouver le meilleur modèle possible dans les circonstances. Dans tous les cas, on aboutit à un compromis dicté par de nombreuses considérations pratiques, comme le budget et le temps disponible, les points essentiels que les clients, les perturbations qu'un modèle d'évaluation peut avoir sur le fonctionnement normal du programme, etc.

Pour décider du modèle à retenir, il faut essentiellement maximiser la crédibilité et l'utilité des conclusions. L'évaluateur doit anticiper les types d'arguments que d'autres invoqueront pour contester les conclusions.

Dans l'évaluation sommative, il faut principalement en arriver à attribuer une cause au résultat du programme, par opposition à d'innombrables autres causes possibles. Dans le langage des évaluateurs, il s'agit de la ***validité interne.*** Toute bonne évaluation sommative est conçue de sorte que soient minimisées les « menaces à la validité interne », c'est-à-dire qu'elle est conçue

de façon à isoler l'effet du programme de l'effet d'autres causes possibles. Campbell et Stanley (1971) ont relevé sept menaces à la validité interne (pièce 1).

Pièce 1 — Menaces à la validité interne

Menaces attribuables à des changements réels dans le contexte ou chez les participants

Historique : Situation où des changements dans le contexte se produisent en même temps que le programme et modifient le comportement des participants (p. ex., une récession pourrait faire mal paraître un bon programme).

Évolution : Situation où des changements se produisent chez les personnes qui participent au programme en raison d'une évolution naturelle biologique ou psychologique.

Menaces attribuables au fait que les participants ne sont pas représentatifs de la population

Sélection : Situation où l'attribution à des groupes participants ou non participants donne des groupes ayant des caractéristiques différentes. Les différences antérieures au programme peuvent être confondues avec celles qui découlent de l'effet du programme.

Mortalité Situation où des participants abandonnent le programme. Ceux qui agissent ainsi peuvent être différents de ceux qui restent.

Régression statistique : Tendance pour les participants qui obtiennent des points très élevés ou très bas dans une mesure de sélection à ne pas obtenir des résultats aussi extrêmes au prochain test. Par exemple, si seuls les participants qui ont eu les plus mauvais résultats dans un test de lecture sont inclus dans un programme d'alphabétisation, ils pourraient mieux réussir au prochain test peu importe le programme, seulement parce que les probabilités d'obtenir d'aussi mauvais résultats la prochaine fois sont faibles.

Menaces engendrées par l'évaluateur

Test : Effets sur des tests postérieurs du fait d'avoir subi un test préalable. On peut mieux réussir au deuxième test tout simplement parce qu'on l'a déjà subi[1].

Instruments : Changement des observateurs, des points obtenus ou de l'instrument de mesure utilisés d'une fois à l'autre.

---

1 Aussi, le participant qui subit un test préalable peut ainsi se sensibiliser à un programme. Il peut avoir de meilleurs résultats tout simplement parce qu'il sait qu'il subit un test, ce qu'on appelle l«l'effet Hawthorne».

4.0 Types de modèles

Il existe des douzaines de modèles possibles qui peuvent servir à déterminer l'effet d'un programme. On trouvera ci-après un bref aperçu des plus répandues.

Comme les modèles sans groupe témoin sont très lacunaires pour ce qui est de la validité interne, c'est-à-dire qu'elles ne peuvent normalement pas exclure d'autres explications aux résultats observés du programme, ***dans toute bonne évaluation sommative, il doit y avoir un groupe témoin.*** Les modèles à un seul groupe sont néanmoins très courantes.

Modèle à un seul groupe

Le modèle d'évaluation le plus simple mais le moins satisfaisant est le *modèle avec test postérieur seulement,* symbolisée par X O (où X représente l'intervention d'un programme, p. ex., un cours, et O représente l'observation postérieure au programme, comme le revenu annuel). Dans ce cas, après que les participants ont terminé le programme en question, on les interroge pour savoir où ils en sont relativement aux comportements ou aux attitudes à modifier. *On ne peut utiliser ce modèle pour attribuer avec fiabilité des effets au programme,* car il n'existe aucun fondement objectif qui permette de supposer que le programme a entraîné des changements. En effet, comme on ne possède aucun renseignement sur la situation des variables évaluées antérieurement au programme, ce modèle ne donne aucune information sur le changement.

Toutes les fois qu'un programme est censé susciter un changement, il faut absolument prendre des mesures avant et après dans le cas du modèle à un seul groupe. La plus simple, soit le *modèle avec test antérieur et test postérieur,* symbolisée ainsi : O_1 X O_2 (où X représente l'intervention et O_1 et O_2 correspondent aux mesures du résultat avant et après le programme), exige un test préalable quelconque avant que le programme n'ait lieu (un test de lecture, par exemple) et un test postérieur au programme. Ce modèle est sujet à la plupart des menaces relatives à la validité interne. Le plus souvent, les participants peuvent avoir changé ou un événement extérieur peut avoir causé une différence observée quelconque entre O_1 et O_2, de sorte qu' *aucun changement ne peut avec crédibilité être attribué au programme.* Par exemple, si une évaluation selon ce modèle montre que le revenu moyen postérieur au programme est plus faible que le revenu moyen antérieur au programme, cela n'est pas nécessairement une indication que le programme n'était pas bon : une récession pourrait avoir entraîné la baisse du revenu.

Dans le modèle avec séries chronologiques, il s'agit de recueillir des données à plusieurs moments au sujet de la situation des participants. En voici la représentation symbolique :
O_1 O_2 O_3 X $O_4 \ldots O_n$

Ce modèle peut servir à exclure (ou du moins à quantifier) la régression et l'évolution en tant que menaces à la validité interne. C'est-à-dire que toutes les tendances personnelles en l'absence du programme peuvent être relevées et gardées fixes. Il faut avoir recours à des

méthodes statistiques avancées pour isoler l'effet du programme. L'historique reste la menace principale. Même s'il peut suffire de disposer de données supplémentaires sur le contexte pour exclure des événements qui peuvent être bien cernés, il est extrêmement difficile de cerner - et encore moins de quantifier - tous les événements possibles qui pourraient avoir entraîné le résultat observé. Voici un exemple. Disons que la Province A a lancé un programme de formation sur une vaste échelle à l'intention de ses bénéficiaires de l'aide sociale, et qu'elle a observé les statistiques sur le nombre de bénéficiaires pendant plusieurs mois avant et après l'intervention pour savoir si le programme diminue la dépendance de l'aide sociale. Or, à peu près au même moment, la Province B a mis en place une nouvelle politique, qui lui est propre, soit la réduction des prestations d'aide sociale pour les clients employables. Il pourrait ainsi y avoir un afflux de bénéficiaires de l'aide sociale de la Province B vers la Province A. Sauf si la Province A était au courant de la nouvelle politique de la Province B et a pris des mesures qui lui permettent d'évaluer l'effet, l'évaluation avec séries chronologiques pourrait sous-estimer tout effet positif du programme de formation.

Bref, le modèle avec un seul groupe rend impossible à toutes fins utiles toute évaluation sommative sérieuse. Elle est notoirement faible et facilement rejetée, parce qu'il est généralement impossible d'exclure d'autres explications possibles, surtout des événements importants qui se sont produits pendant que le groupe expérimental participait au programme (p. ex., une récession). Il faut recourir à le modèle à deux groupes pour faire une évaluation sérieuse.

Modèle à deux groupes

Pour déterminer de façon valable l'effet, il faut comparer les résultats obtenus par un groupe de personnes qui ont participé au programme (groupe expérimental) avec un groupe équivalant de personnes qui n'y ont pas participé (groupe témoin). En théorie, la meilleure façon de faire cela consiste à recourir à une *expérience par randomisation*, selon laquelle les personnes sont dirigées au hasard vers le groupe expérimental ou le groupe témoin. (Rossi et Freeman, 1993) Les mesures des résultats, choisies en fonction des objectifs du programme, sont observées à un intervalle donné après la fin de l'intervention, ainsi que les différences entre les groupes qui sont attribuables au programme, c'est-à-dire qu'on peut affirmer que le programme a causé les différences observées. Le modèle est représenté comme suit :

X O [participants]
O [non-participants]

Comme la randomisation devrait servir à éliminer, du moins en moyenne, toutes les différences systématiques qui existent entre les groupes, il n'est pas nécessaire d'administrer un test préalable. On peut tout simplement mesurer l'effet du traitement en comparant les résultats moyens des groupes expérimentaux et des groupes témoins, et expliquer les différences attribuables au hasard par les techniques statistiques courantes (Greenberg et Wiseman, 1992).

La menace principale associée à ce modèle - en supposant que la randomisation ait été bien faite - est la mortalité ou l'attrition non attribuable au hasard (c.-à-d. que des membres du groupe participant abandonnent avant la fin du programme, ou que des membres du groupe participant et du groupe témoin ne peuvent être repérés pour le suivi pour des raisons qui ne

sont pas attribuables au hasard et qui, par conséquent, peuvent être systématiquement liées aux effets du programme). Pour cette raison, un test préalable est souvent administré au groupe expérimental et au groupe témoin (Mark et Cook, 1984, jugent que cela est « essentiel »), de sorte que les effets de l'abandon du programme peuvent être quantifiés et pris en compte dans l'analyse.

Même si les modèles expérimentaux sont en théorie le plus près possible de l'idéal, ils sont rarement pratiques. Les contraintes de loin les plus courantes tiennent aux employés du programme, qui refusent de se conformer parce qu'ils jugent que la sélection randomisée pose des problèmes d'ordre éthiques. En outre, souvent, l'évaluateur entre en scène longtemps après que l'attribution aléatoire aurait dû avoir lieu. Il existe également d'autres problèmes. Le plus sérieux, c'est que les méthodes expérimentales ne servent habituellement qu'à déterminer l'effet moyen du programme; elles ne peuvent répondre à de nombreuses questions d'ordre stratégique, notamment l'effet médian du programme et la proportion de participants ayant subi un effet positif (ou négatif) en raison du programme (Heckman et Smith, 1995)[2] .

En raison de ces contraintes, les modèles quasi expérimentaux (non expérimentaux) sont souvent les seuls qui soient satisfaisants. Il y a différents *modèles quasi expérimentaux*, mais la méthode la plus courante et la plus sûre consiste à former un groupe témoin de personnes qui présentent un profil comparable à celui des participants. On peut y arriver soit en tenant compte statistiquement de l'effet des différences entre les groupes pendant l'analyse des données, soit en appariant les participants et les non-participants selon des traits caractéristiques (comme l'âge, le sexe et la scolarité) qui pourraient influer sur les résultats attendus, soit les deux[3] . Il s'agit d'en arriver à une attribution aléatoire la plus authentique possible en essayant de minimiser ou de tenir compte de l'effet des différences entre les groupes. Voici la représentation[4] :

O_1 X O_2 [participants]

O_1 O_2 [non-participants]

---

[2] On trouvera à l'annexe A un bref exposé du débat en cours sur le choix entre les méthodes expérimentales et quasi expérimentales.

[3] Rubin (1979) a montré que l'utilisation des deux techniques – appariement et rajustement statistique - était préférable à l'utilisation d'une seule des deux techniques (dans Dickinson et coll., 1987).

[4] À noter que l'information préalable au programme n'est pas strictement nésessaire pour certaines modèles avec groupe témoin - voir la partie ci-après sur la façon de prévenir le biais de sélection - mais elle est toujours souhaitable.

Dans ce cas, X représente l'intervention du programme, O_1 est une observation antérieure au programme et O_2 est une observation postérieure au programme. Par exemple, O_1 pourrait être le revenu annuel en 1995, O_2 le revenu annuel en 1997 et X un programme de formation en 1996.

Avec un modèle quasi expérimentale longitudinale, par exemple, l'évaluateur compare les résultats des deux groupes : les participants au programme (le « groupe expérimental ») et les non-participants (le « groupe témoin »)[5] . Les « résultats », qui se rapportent aux objectifs du programme de formation - trouver un emploi, par exemple, - sont habituellement déterminés par le truchement d'une enquête de suivi, menée des mois ou même des années après que le participant a terminé le programme. On compare le résultat postérieur au programme pour chaque groupe à la situation antérieure au programme pour déterminer s'il y a eu, en moyenne, un changement à l'intérieur de chaque groupe. On utilise ensuite des méthodes statistiques courantes pour déterminer si le changement est significativement différent entre les groupes.

Disons, par exemple, qu'une enquête de base (voir ci-après) a montré que la moitié des participants et la moitié des non-participants ont travaillé une année avant de demander l'assurance-emploi, et que, dans une enquête de suivi, on a constaté que 70 % des personnes qui ont reçu une formation travaillaient un an après avoir reçu cette formation, par opposition à 60 % du groupe témoin un an après qu'il eut cessé de toucher des prestations d'assurance-emploi. L'augmentation pour ceux qui ont reçu la formation est par conséquent de 20 %, par opposition à 10 % pour les autres. Il suffit de simples tests statistiques pour déterminer si cet écart est significatif.

Mais ce n'est pas parce que l'on constate un écart statistiquement significatif entre des groupes qu'il faut attribuer cet écart au programme. *L'analyste doit démontrer que l'écart est attribuable au programme*, c'est-à-dire qu'il lui faut exclure les menaces à la validité interne. Malheureusement, les données empiriques montrent que les participants sont susceptibles d'être différents des non-participants selon des particularités qui ont un effet sur les variables des résultats. La sélection dans la plupart des programmes ne se fait pas de façon aléatoire : ceux qui choisissent de participer peuvent être plus motivés que ceux qui ne le font pas, par exemple; et les administrateurs du programme choisissent plus souvent qu'autrement les personnes qui, à leur avis, ont les meilleures chances de réussite (c.-à-d. les plus talentueux) ou, inversement, celles qui ont le plus besoin de l'intervention.

Peu importe sa source, le *biais de sélection* a un effet sur la comparabilité des groupes expérimentaux et témoins. Tant que toutes les différences entre les groupes comparés sont observables (p. ex., les caractéristiques personnelles), le biais de sélection ne constitue pas un obstacle parce que des méthodes statistiques, comme l'analyse de régression multiple, peuvent tenir compte de l'effet des écarts. Les chercheurs font tout ce qu'ils peuvent pour apparier les

---

[5] On trouvera à l'annexe B une représentation mathématique élémentaire de la méthode quasi expérimentale.

individus qui subissent l'intervention et les échantillons témoins pour s'assurer que les caractéristiques observées sont très semblables, mais ils savent très rarement pourquoi une personne participe à un programme. Si une caractéristique inconnue (et, par conséquent, dont on ne tient pas compte de l'effet) de la personne ou du programme a influé sur la décision de participation, la sélection est alors non aléatoire et les écarts entre les participants et les non-participants peuvent être à tort attribués au programme.

Il n'existe aucune méthode statistique qui puisse complètement résoudre le problème du biais de sélection. Comme il est impossible d'anticiper tous les facteurs qui ont joué dans la décision de participation, il est impossible de concevoir les enquêtes et les protocoles devant servir à recueillir tous les renseignements utiles. Pour mener une quasi expérience, il faut avoir recours à des techniques d'analyse qui sont beaucoup plus compliquées que celles des expériences authentiques. Il faut avoir recours à des statistiques de haut niveau - à des « modèles économétriques » - pour tenir compte des écarts entre les groupes et isoler l'effet du programme.

5.0 Exemples de biais de sélection

Il existe un certain nombre de méthodes qui permettent de contourner le problème possible du biais de sélection. Pour les décrire, il est utile de donner certains exemples concrets des formes que le biais de sélection prend fréquemment dans les études d'évaluation sur les programmes d'éducation ou de formation.

Premier exemple : programme d'alternance travail-études et autres types de programmes d'études

Dans un programme d'alternances travail-études, l'étudiant alterne systématiquement entre des périodes d'études et des périodes de travail, tandis que dans les autres types de programmes, il suit tout simplement des cours. De nombreux observateurs estiment que la combinaison des études et des expériences de travail est susceptible d'améliorer l'employabilité et le revenu des diplômés des programmes d'alternance travail-études. Existe-t-il des données empiriques soutenant cette affirmation ? Pour répondre de façon simple à cette question, on pourrait comparer la situation d'emploi et de revenu d'un échantillon de diplômés des programmes d'alternance travail-études et à celle d'autres types de programmes d'études. Obtiendrait-on une preuve manifeste de l'effet des programmes d'alternance travail-études sur l'employabilité et le revenu ? La réponse est généralement est non.

Il peut y avoir des raisons pour lesquelles les personnes qui ont suivi un programme d'alternance travail-études ne se retrouvent pas dans la même situation d'employabilité et de revenu que celles qui suivent d'autres types de programmes, et que ces raisons n'ont rien à voir avec les programmes d'études comme tels. Par exemple, s'il y a un nombre limité de programmes d'alternance travail-études (comme c'est généralement le cas), il se peut que ces programmes admettent des étudiants plus aptes en moyenne que ne le font les autres programmes comparables. Aussi, les étudiants qui demandent de participer à un programme d'alternance travail-études sont peut-être en moyenne plus déterminés à embrasser une carrière que ceux qui s'inscrivent à d'autres types de programmes. Pour ces deux raisons, les diplômés des

programmes d'alternance travail-études pourraient présenter un niveau d'employabilité et de revenu plus élevé, même s'ils ne suivaient pas un tel programme d'alternance travail-études. C'est que, si les diplômés d'un tel programme avaient plutôt suivi un autre type de programme, leur employabilité et leur revenu auraient été plus élevés que ceux des diplômés d'autres types de programmes. Dans une certaine mesure, les diplômés des programmes d'alternance travail-études présentent un niveau plus élevé d'employabilité et de revenu parce qu'ils sont de meilleurs étudiants, plus aptes, et parce qu'ils sont davantage déterminés à faire carrière que leurs homologues d'autres programmes.

Bien sûr, il est également possible que les programmes d'alternance travail-études aient un effet favorable sur l'employabilité et sur le revenu des diplômés. Si tel est le cas, la différence totale observée entre les diplômés de ces programmes et les autres se répartit en deux volets : dans un premier temps, parce que les programmes d'alternance travail-études attirent des étudiants plus aptes et plus déterminés à faire carrière (l'effet de sélection), et dans un deuxième temps, en raison de l'effet du programme.

Voilà le problème du biais de sélection possible. Les responsales des programmes d'alternance travail-études choisissent les étudiants les plus aptes et les plus déterminés à faire carrière, dont l'employabilité aurait été plus élevée même s'ils n'avaient pas suivi le programme. Dans ce cas, la simple comparaison des diplômés des programmes d'alternance travail-études et des autres types de programmes donnerait lieu à une surestimation de l'effet véritable du programme. Toutefois, comme nous en parlons ci-après, le biais de sélection peut être soit positif, soit négatif, c'est-à-dire que, si on n'en tient pas compte, l'effet estimatif peut être supérieur ou inférieur à l'effet véritable.

2e exemple : Écoles privées et écoles publiques

Bon nombre des mêmes questions se posent si l'on veut comparer les résultats (comme les résultats moyens obtenus dans les examens normalisés, les taux d'achèvement des études secondaires ou la réussite des études postsecondaires) dans les écoles privées et publiques. Les écoles privées choisissent peut-être des élèves plus aptes en moyenne que leurs contreparties publiques. De même, l'élève moyen qui fréquente l'école privée peut être plus susceptible de présenter d'autres caractéristiques (ses parents attachent plus d'importance à l'éducation) que l'élève moyen qui fréquente l'école publique.

Pour ces raisons, il est peu probable qu'une simple comparaison des résultats des élèves des écoles privées et des écoles publiques puissent aboutir à des estimations non biaisées de l'effet des écoles privées sur ces résultats. Certaines des différences observées existent parce que les écoles privées et les écoles publiques choisissent des élèves qui sont systématiquement différents relativement à des caractéristiques comme les aptitudes, et l'attitude de la famille envers l'éducation.

3e exemple : Effet des programmes de formation parrainés par l'État

Comme on pourra maintenant s'en rendre compte, des problèmes semblables de sélection se posent lorsqu'il s'agit d'évaluer les effets des programmes de formation. Ceux qui suivent une formation sont susceptibles d'être systématiquement différents de ceux qui ne le font pas. Cela peut être attribuable à des différences entre les personnes qui suivent la formation et celles qui ne le font pas : par exemple, les personnes qui sont plus scolarisées ou qui veulent davantage intégrer le marché du travail peuvent être plus susceptibles de demander à suivre une formation. En revanche, ces différences peuvent se manifester en raison de la sélection par les administrateurs du programme, qui peuvent être plus enclins à retenir les personnes qui sont le plus aptes à profiter de la formation.

Le présent exemple illustre aussi maintenant pourquoi le biais de sélection peut être soit positif, soit négatif. Supposons qu'un programme soit conçu pour aider les plus désavantagés dans une population donnée. Dans ce cas, il est probable qu'une simple comparaison de ceux qui suivent le programme et de ceux qui ne le font pas sous-estimerait l'effet véritable du programme.

Ces trois exemples montrent que, à toutes fins utiles, toute évaluation de programme est susceptible de se heurter à un biais de sélection. C'est que, dans presque tous les programmes ou interventions, les participants et les non-participants ont des choix à faire, tout comme les personnes qui administrent le programme. En raison de ces choix, les groupes de participants et de non-participants sont susceptibles de différer sous des aspects systématiques (ou non aléatoires). Si ces différences entre les participants et les non-participants sont également associées aux résultats du programme, la simple comparaison entre les participants et les non-participants aboutit à des estimations biaisées sur l'effet du programme. Pour corriger ce problème très courant, il existe un certain nombre de méthodes statistiques.

6.0 Méthodes de correction du biais de sélection

Avant de décrire les diverses méthodes qui ont été élaborées et qui servent largement à atténuer le problème du biais de sélection possible, nous voulons faire deux observations préliminaires. En premier lieu, comme on peut le voir dans les exemples ci-dessus, la sélection dans un programme peut être fondée sur des facteurs observables ou des facteurs non observables. Par exemple, dans le cas de la comparaison des programmes d'alternance travail-études et autres, on peut observer les aptitudes des étudiants du programme d'alternance travail-études et des autres (p. ex., leurs diplômes d'études secondaires), mais on ne peut savoir dans quelle mesure les étudiants ont le désir de faire carrière. C'est la richesse des données disponibles qui détermine quels facteurs sont observés et lesquels ne le sont pas.

Il est facile de tenir compte de l'effet de la sélection en vertu du programme, laquelle est faite selon des facteurs observables. Par conséquent, plus les données disponibles sont riches (et moins il y a de facteurs inobservables), plus l'envergure du biais de sélection attribuable à des facteurs inobservables est réduite.

La deuxième observation est la suivante : bien qu'il existe certains facteurs inobservés qui influent sur la sélection pour le programme (ce qui est presque toujours le cas), ceci n'implique pas nécessairement que les simples comparaisons des participants et des non-participants seront assujetties au biais de sélection. Il y a un biais de sélection lorsque les facteurs inobservés qui influent sur la participation ou la non-participation au programme influent également sur les effets du programme.

Pour clarifier ce point, prenons un exemple extrême. Supposons que nous voulons comparer les écoles publiques et les écoles privées sous l'angle des résultats scolaires, comme la performance des élèves dans des test normalisés. Supposons qu'en moyenne les écoles privées reçoivent plus d'élèves avec un jour de naissance pair que ne le font les écoles publiques. Pour le chercheur, la caractéristique qui consiste à « avoir pour jour de naissance un nombre pair ou impair » n'est pas observée. Il s'agit d'un cas de sélection non aléatoire : si les élèves avaient été choisis au hasard dans les écoles publiques et les écoles privées, la proportion d'élèves avec un jour de naissance pair serait environ égale dans les deux types d'école. Toutefois, dans la mesure où le fait d'avoir un jour de naissance pair n'a pas d'effet sur les résultats évalués (la performance des élèves dans des tests normalisés), cette sélection non aléatoire n'aura pas un effet de distorsion sur une simple comparaison des résultats des élèves des écoles publiques et des écoles privées.

Pour ces raisons, les méthodes de correction du biais de sélection portent surtout sur les problèmes possibles associés aux facteurs qui influent sur la sélection pour le programme qui sont i) non observés par le chercheur ou l'évaluateur et ii) sont en corrélation avec les résultats évalués du programme. Malheureusement, dans plusieurs études d'évaluation, il y a un grand nombre de ces facteurs qui répondent à ces deux conditions. Pour cette raison, nous recommandons fortement que le biais de sélection possible soit pris en compte dans toute étude d'évaluation.

De par leur nature même, les facteurs qui donnent lieu au biais de sélection ne sont pas observés. Dans certains cas, les méthodes décrites ci-après amèneront l'évaluateur à conclure que de telles sources possibles de biais de sélection ne sont pas quantitativement importantes et, par conséquent, n'engendrent pas un biais. C'est peut-être du fait que les facteurs non observés conduisant à une sélection non aléatoire dans le programme ne sont pas quantitativement importants dans le programme en question, ou encore parce que ces facteurs sont quantitativement importants mais ne sont pas en corrélation avec les résultats du programme à étudier. Dans d'autres cas, les méthodes décrites ci-après amèneront l'évaluateur à conclure que le biais de sélection est quantitativement important. Dans ces cas, les méthodes procurent également une idée estimative de l'ampleur du biais de sorte qu'on puisse en dériver une estimation des effets véritables du programme sur les résultats.

6.1 Méthodes de correction en deux étapes

Des méthodes de correction en deux étapes relativement au biais de sélection ont été élaborées par James Heckman et d'autres à la fin des années 1970, et sont devenues les méthodes les plus répandues. Dans un premier temps, la probabilité de participation au programme est

analysée. Cette analyse consiste habituellement en un modèle d'équation simple dans lequel la variable dépendante est la probabilité de participation au programme (une variable indicatrice qui est égale à l'unité pour les participants au programme et à zéro pour les non-participants) et les variables indépendantes sont des facteurs divers qui auraient un effet sur la participation ou la non-participation au programme. Il s'agit essentiellement dans la première étape d'obtenir un facteur de correction (appelé « inverse du rapport de Mill ») qui sert dans la deuxième étape à tenir compte d'un biais de sélection possible. Aussi, les estimations obtenues dans la première étape peuvent être d'intérêt en elles-mêmes en ce sens qu'elles donnent une idée de l'importance des divers facteurs qui influent sur la participation ou la non-participation au programme.

La deuxième étape consiste à évaluer l'effet du programme au moyen d'un modèle spécifique (une équation). Le modèle comprend :

- une « variable dépendante », soit le résultat sur lequel le programme de formation est censé avoir un effet, par exemple, le revenu;
- plusieurs variables « indépendantes » ou explicatives, qui sont les facteurs observés censés avoir un effet sur le résultat (p. ex., l'âge, le sexe, la scolarité);
- la variable de la « correction du biais de sélection » (ou inverse du rapport de Mill) obtenue dans la première étape;
- une variable indicatrice relative à la participation ou à la non-participation au programme;
- un terme d'erreur aléatoire pour tenir compte des forces non observées qui pourraient influer sur la mesure des résultats.

Voici le modèle en mots :

> Revenu = effet des divers facteurs observés + effet du biais de sélection + effet du programme + erreur aléatoire

(Voir l'annexe C pour l'équation mathématique et d'autres explications.)

Ainsi, le modèle isole l'effet du programme d'autres facteurs d'influence possibles. Si le modèle est bien défini, l'ajout de la variable de la « correction du biais de sélection » élimine ce biais possible, de sorte qu'on obtient des estimations non biaisées de l'effet du programme. Nous reviendrons ci-après à la question importante de la façon de déterminer si le modèle est bien défini.

Il existe un moyen utile d'interpréter cette méthode en deux étapes. L'on sait très bien que, si une variable importante est omise dans un modèle, on obtient des estimations biaisées des coefficients sur les variables comprises dans le modèle. Faute d'une méthode qui permette de tenir compte de la sélection dans le programme, l'estimation de l'équation du résultat omet un facteur important : les facteurs déterminants de la participation au programme. Le terme de la

« correction du biais de sélection » obtenu dans la première étape donne une idée estimative de ce facteur. C'est pourquoi l'inclusion de ce terme permet l'obtention d'estimations non biaisées (si le modèle est bien défini).

Il convient de faire une dernière observation au sujet de cette méthode à deux étapes. Il est important d'avoir une ou plusieurs variables qui influent sur la sélection dans le programme (c.-à-d. qui entre dans l'équation de la première étape) mais qui n'influent pas sur les résultats du programme (c.-à-d. qui n'entre pas dans l'équation de la deuxième étape). Avec de telles variables, il est possible de distinguer la participation au programme, d'une part, et les résultats où les effets du programme, d'autre part. Outre cette « variable différenciatrice » qu'il est important d'avoir, l'équation de la participation à la première étape et l'équation des résultats à la deuxième étape peuvent avoir de nombreuses variables en commun. L'importance de ces « variables différenciatrices » apparaît également dans le contexte de la méthode exposée ci-après.

6.2 Méthodes des variables instrumentales

Il existe un biais de sélection en raison de la corrélation entre la variable indicatrice de participation ou de non-participation au programme et le terme d'erreur aléatoire dans l'équation du résultat. La méthode des « variables instrumentales » (VI) pour la résolution du problème du biais de sélection, abordée notamment par Heckman et Robb (1985) et par Moffitt (1991), consiste essentiellement en la recherche d'une variable (ou de variables) qui influe sur la sélection dans le cadre du programme, mais non sur le résultat du programme (et n'est par conséquent pas en corrélation avec le terme d'erreur aléatoire dans l'équation du résultat). Comme la variable instrumentale n'est pas en corrélation avec le terme d'erreur aléatoire, elle peut servir à l'estimation sans qu'il y ait un biais. On trouve à l'annexe C le modèle pour l'estimateur de la variable instrumentale.

La recherche des variables instrumentales exige l'examen en profondeur [illegible]rocessus de sélection. Les caractéristiques personnelles des individus sont rarement suffisantes en tant que variables instrumentales parce qu'elles sont habituellement liées au résultat. Par exemple, le niveau de scolarité a probablement un effet sur l'employabilité. Moffit laisse entendre que si la disponibilité de l'intervention varie, cette variable peut être pertinente. Si un programme de formation est disponible dans une région mais non dans une autre pour des raisons non reliées aux résultats anticipés du programme, la région est une variable instrumentale légitime. Cela peut être le cas si le programme est offert pour des raisons autres que politiques, bureaucratiques ou économiques.

Pour qu'elle soit un « instrument » légitime, la variable doit être liée à la participation ou à la non-participation au programme, mais non aux résultats du programme. Dans certaines situations, il peut y avoir de nombreuses variables instrumentales possibles. Dans ces cas, comment l'analyste doit-il choisir parmi toutes celles qui sont possibles ? Voici la réponse à cette question. Chaque VI qui n'est manifestement pas liée au résultat du programme (c.-à-d. qui n'est pas en corrélation avec le terme d'erreur aléatoire dans l'équation du résultat) donne des estimations non biaisées de l'effet du programme. Toutefois, certaines VI peuvent donner

des estimations plus précises de l'effet du programme. Plus précisément, plus la VI est en corrélation avec la participation ou la non-participation au programme, plus sont précises les estimations de l'effet du programme. Par conséquent, la difficulté qui se pose dans l'estimation de la VI consiste à trouver une telle variable qui est en forte corrélation avec la participation au programme, mais non avec le résultat du programme. Il est malheureusement souvent difficile de trouver des variables qui répondent à ces deux exigences et, par conséquent, difficile de trouver de bonnes VI parmi toutes celles qui sont possibles.

Pour mieux comprendre les principes de l'estimation des VI, prenons le cas d'un programme dans lequel la participation ou la non-participation est déterminée par attribution aléatoire. Supposons que chaque candidat est considéré comme un participant P (non-participant NP) si, en jouant à pile ou face, on obtient Pile (Face). Alors, la variable indicatrice Pi (qui est égale à l'unité si on obtient Pile et à zéro si on obtient Face) est une variable instrumentale idéale parce que Pi est en parfaite corrélation avec la variable indicatrice de la participation et parce que Pi n'est pas en corrélation avec le résultat du programme. En pratique, bien sûr, il est rare que l'on dispose d'une telle VI idéale, mais cet exemple illustre les caractéristiques qu'on cherche à obtenir lorsqu'on utilise cette méthode.

Cette méthode a donc certaines caractéristiques en commun avec les « variables différenciatrices » utilisées dans la méthode en deux étapes exposée ci-dessus. Dans les deux cas, en effet, il faut des données semblables. Les différences principales sont les suivantes : i) la méthode de la variable instrumentale s'effectue en une seule étape et n'exige par conséquent pas la modélisation explicite du processus de participation au programme; et (ii) la méthode de la variable instrumentale engendre des estimations sans biais de sélection (si le modèle est bien défini), mais elle ne procure pas une estimation de l'ampleur du biais de sélection comme dans le cas de la méthode en deux étapes[6] .

On pourrait donner de nombreux exemples de « variables instrumentales ». Moffitt (1991) fait mention d'un programme de consultation en santé qui est financé par l'État et qui, pour des raisons non liées aux besoins de santé des populations des deux secteurs, existe dans un secteur de la ville mais non dans l'autre. C'est ainsi qu'une variable indicatrice pour les deux secteurs de la ville n'est pas liée aux besoins de santé de la population dans les deux secteurs, mais influe sur la participation au programme. Un autre exemple, lié à l'évaluation des écoles publiques face aux écoles privées, dont il a été question ci-dessus, serait une mesure de la proximité d'une école privée pour chacun des élèves de l'échantillon. Cette proximité pourrait influer sur la probabilité de fréquentation d'une école privée, mais non sur l'effet de cette fréquentation sur les résultats de l'élève. On trouvera à l'annexe C de plus amples renseignements sur la méthode de VI, y compris les équations.

---

[6] En outre, la méthode de l'estimation des VI n'exige pas l'hypothèse des termes d'erreur aléatoire normalement distribués dans l'équation selon la méthode des probits de la première étape utilisée dans la méthode en deux étapes.

6.3 Méthodes longitudinales

La méthode en deux étapes et la méthode des variables instrumentales expliquées ci-dessus peuvent être appliquées avec des données postérieures au programme seulement (c.-à-d. des données transversales sur les participants et sur les non-participants). Toutefois, si des données sur les participants et les non-participants antérieures au programme sont disponibles, on peut également les utiliser; si elles sont intégrées à ces méthodes, on obtiendra généralement des estimations plus précises et plus crédibles de l'effet du programme. Avec les données longitudinales, on suit les mêmes individus sur deux périodes ou plus, et les méthodes longitudinales exposées ci-après exigent au moins une observation antérieure au programme et au moins une observation postérieure au programme, tant sur les participants que sur les non-participants.

L'estimateur longitudinal le plus courant de l'effet d'un programme est celui des « effets fixes » ou « de la différence à l'intérieur des différences » (parfois appelé tout simplement « estimateur des différences »). Dans le cas le plus simple, où il y a une seule observation antérieure au programme et une seule postérieure au programme, on procède comme suit. En premier lieu, on détermine la différence qui existe entre la valeur postérieure au programme de la mesure du résultat et la valeur antérieure au programme de la mesure du résultat pour chaque participant et chaque non-participant. Cette différence est, par conséquent, une mesure du changement qui a été observée dans le résultat entre la période antérieure au programme et la période suivant le programme. Si le résultat évalué est le revenu, comme c'est le cas dans l'évaluation des programmes de formation, il s'agirait du gain ou de la perte de revenu pour chaque participant et non-participant. Ensuite, on détermine la différence qui existe entre le changement moyen antérieur au programme par rapport au changement moyen postérieur au programme pour les participants et le changement moyen antérieur au programme par rapport au changement postérieur au programme pour les non-participants (voir l'équation à l'annexe C). Dans le cas d'un programme de formation, il s'agit tout simplement de la différence entre le gain moyen ou la perte moyenne de revenu des participants, et le gain moyen ou la perte moyenne des non-participants.

Cet estimateur simple de l'effet du programme n'aura pas de biais de sélection si la sélection pour le programme dépend d' «effets fixes » non observés propres à la personne, c'est-à-dire de facteurs qui sont propres (ou particuliers) à chaque individu dans l'échantillon, mais qui sont constants (« fixes ») sur la période visée par l'analyse. Par exemple, dans la comparaison des programmes d'alternance travail-études et autres types de programmes, un tel facteur non observé pourrait être le désir, chez les étudiants, de faire carrière. Si, comme nous l'avons expliqué précédemment, la sélection dans les programmes d'alternance travail-études et d'autres types de programmes est fonction du désir, chez l'étudiant, de faire carrière, et si ce facteur n'est pas observé par le chercheur (comme c'est souvent le cas), le biais qui existerait en raison de ce facteur non observé est éliminé grâce à l'estimateur de la différence à l'intérieur des différences pourvu que le désir, chez l'étudiant, de faire carrière, soit constant pendant la période visée. De même, dans le cas des programmes de formation, l'hypothèse des « effets fixes » doit être retenue si des facteurs non observés propres à la personne, comme l'ambition, l'activité sur le marché du travail et l'aptitude à la formation, sont constants (mais peuvent

varier selon les individus). Dans de nombreux cas, une telle hypothèse d' «effets fixes » pourra sembler raisonnable pour de tels facteurs non observés propres à la personne, même si la validité de l'hypothèse doit être vérifiée, comme il en est question ci-après.

En résumé, l'estimateur de la différence à l'intérieur des différences donne lieu à des estimations non biaisées de l'effet du programme - même en présence d'un biais de sélection possible - lorsque la source du biais possible est une corrélation entre la participation ou la non-participation au programme et un facteur non observé, qui peut être différent d'un individu à un autre, mais qui est constant dans le temps pour chaque individu. Si la sélection dans le cadre du programme prend cette forme, le simple estimateur de la différence à l'intérieur des différences est une façon simple de tenir compte de l'effet du biais de sélection.

Des estimateurs longitudinaux plus compliqués existent pour les cas où l'hypothèse des effets constants ou « fixes » propres à la personne ne convient pas. Il faut généralement dans ce cas plus d'une observation antérieure au programme et une autre postérieure au programme pour chaque individu. Par exemple, Moffitt (1991) parle d'un estimateur de la « différence à l'intérieur des différences dans les taux de croissance », lequel convient lorsque le changement d'une période à l'autre dans l'effet propre à la personne est constant dans le temps. Cet estimateur exige au moins deux observations antérieures au programme et deux autres qui en sont postérieures. Ashenfelter et Card (1985) expliquent d'autres types d'estimateurs longitudinaux dans le contexte des programmes de formation.

6.4 Tests de caractérisation d'autres modèles

Nous avons exposé trois catégories générales de méthodes servant à tenir compte de l'effet du biais de sélection. Pour chacune de ces catégories générales, il existe un certain nombre de variantes de la méthode de base. La question se pose naturellement : laquelle de ces méthodes employer et dans quelles circonstances ?

La réponse à cette question importante est, en partie, qu'il appartient au chercheur / évaluateur de choisir la méthode qui convient, selon les particularités du programme à évaluer. Le bon évaluateur est notamment en mesure de bien déterminer le modèle, ce qui comprend les méthodes qui permettent de tenir compte de l'effet du biais de sélection possible. Il lui faut, à cette fin, du jugement, de l'expérience et la capacité d'obtenir les renseignements nécessaires sur la nature des modalités selon lesquelles les participants sont choisis pour le programme.

Même si des facteurs tels le jugement et l'expérience sont importants, dans la plupart des cas, ils sont probablement insuffisants pour que l'évaluateur puisse déterminer avec une certitude raisonnable quelle méthode convient davantage pour tenir compte de l'effet du biais de sélection. Pour cette raison, il est important d'avoir recours à toute une gamme de tests de caractérisation pouvant servir à déterminer quels modèles ou caractéristiques conviennent aux données et lesquels ne conviennent pas. On peut trouver dans les textes de Heckman et Hotz (1989) et de Moffitt (1991) des exemples de tels tests de caractérisation. Heckman et Hotz (1989) constatent, dans le cas des programmes de formation, que l'utilisation d'un certain nombre de tests de caractérisation leur permet de diminuer sensiblement le nombre de formes différentes

possibles que la sélection dans le programme peut prendre, ce qui leur permet de réduire sensiblement la fourchette des estimations possibles de l'effet du programme.

À ce jour, de tels tests de caractérisation servant à évaluer la validité d'autres modèles n'ont pas été utilisés autant qu'ils auraient dû l'être. Toutefois, en partie en raison des contributions récentes à la littérature sur l'évaluation, l'utilisation de ces tests devient un aspect de plus en plus important dans la réussite d'évaluations. Il existe un effet secondaire important de cette tendance, soit que les évaluateurs sont de plus en plus tenus de réfléchir et de porter attention au mécanisme selon lequel la sélection pour le programme se fait (et, par conséquent, à la meilleure façon de tenir compte du biais de sélection possible), plutôt que de s'en remettre à une technique mécanique comme la méthode en deux étapes de Heckman.

6.5 Facteurs déterminants de la participation au programme

Une dernière observation : plus on peut obtenir de renseignements au sujet des modalités selon lesquelles les participants complètent le programme et que les non-participants l'abandonnent, plus seront crédibles les estimations de l'effet du programme. Comme nous l'avons dit précédemment, la participation au programme dépend de facteurs à la fois observés et non observés et il est beaucoup plus simple et moins incertain de rendre compte de l'influence des facteurs observés que du rôle des facteurs non observés. Par conséquent, l'acquisition de données plus riches sur les facteurs déterminants de la participation au programme est l'une des méthodes les plus efficaces de tenir compte de l'effet du biais de sélection possible.

Dans toute évaluation, il existera toujours certains facteurs qui ne seront pas observés mais qui pourraient entraîner un biais de sélection. Pour tenir compte le mieux possible de telles possibilités, plus l'évaluateur dispose de renseignements qualitatifs riches sur le programme ainsi que sur les caractéristiques des participants et des non-participants, plus il est en mesure de choisir les meilleures méthodes qui lui permettent de tenir compte de l'effet du biais de sélection.

7.0 Choix des mesures d'impacts

L'une des étapes cruciales de l'évaluation sommative consiste à choisir les meilleures mesures des résultats. Une mesure inadéquate ou non fiable peut complètement annuler la valeur d'une évaluation de l'effet en produisant des estimations trompeuses. (Rossi et Freeman, 1993, p. 234)

Les mesures des résultats se rapportent à l'effet que le programme est censé avoir, de sorte que les bonnes mesures sont liées aux objectifs du programme. En général, les programmes de formation de DRHC comportent la totalité ou un sous-ensemble des buts suivants : rehausser la scolarité; favoriser la transition vers le marché du travail en améliorant l'employabilité et le revenu; réduire la dépendance d'un soutien du revenu passif; améliorer les attitudes face au travail. Ces résultats sont facilement quantifiables et seraient considérés comme des dépendantes variables dans les modèles économétriques (puisqu'il est raisonnable d'évaluer le programme

selon ses effets prévus). Par conséquent, parmi les variables de bons résultats (postérieures au programme) pour les programmes de formation, mentionnons le niveau de scolarité, la situation d'activité (c.-à-d. le fait de travailler ou non), le temps passé au travail ou aux études, le revenu annuel, le nombre de mois où une personne a touché de l'aide sociale, le nombre de semaines où elle a touché des prestations d'a.-e. ainsi que les attitudes face au travail, aux études et à l'aide passive.

Il est important de signaler que des mesures antérieures au programme des mesures des résultats sont fortement souhaitables dans toute évaluation quasi expérimentale. Il est plutôt rare que le système d'information de gestion du programme contienne des données satisfaisantes antérieures au programme pour les participants et les non-participants. Par exemple, DRHC a, sur l'utilisation de l'a.-e., des données complètes et exactes qui sont très utiles pour l'évaluation des effets. Le Système d'évaluation du service et des résultats (SESR) de DRHC dispose aussi de bonnes données longitudinales sur la situation d'activité et sur le temps passé aux études ou au travail. Malheureusement, il n'existe que très peu de données fiables antérieures au programme sur l'utilisation de l'aide sociale et sur les attitudes. Il faudrait néanmoins recueillir de tels renseignements si la chose est possible pour faire une évaluation des effets. On peut recueillir certaines données antérieures au programme lors d'une enquête de suivi, mais ces données peuvent être imprécises parce que les souvenirs sont souvent inexacts et que des documents ont été perdus. D'autres renseignements antérieurs au programme, particulièrement ceux qui portent sur les attitudes, sont impossibles à récupérer après le programme.

La meilleure solution consiste à recueillir un échantillonnage des données antérieures au programme par le truchement d'une « enquête de base ». Il s'agit essentiellement d'établir les caractéristiques antérieures au programme des participants et des non-participants pour faciliter une évaluation sommative plus tard. Dans une bonne enquête de base, on pose des questions qui visent à établir ce que la personne faisait en regard des variables des résultats (travail, études, revenu, aide sociale, a.-e., etc.) avant le programme. Il faut tenir compte de ce qui s'est passé avant le programme chaque année pour au moins deux ans avant le début du programme : par exemple, le nombre de mois où la personne a été bénéficiaire de l'aide sociale en 1995, 1996 et 1997. De plus, une section du questionnaire devrait être consacrée aux attitudes et une autre aux facteurs démographiques (surtout si le système administratif n'est pas fiable ou s'il n'en existe aucun). Enfin, dans l'enquête de base, il faut demander le nom d'au moins deux personnes avec qui communiquer (membre de la famille ou amis) pour retrouver la personne à des fins de suivi, parce que le groupe cible des programmes de formation est souvent très mobile.

8.0 Choix d'un groupe témoin

Au chapitre de la mesure des résultats, l'objectif le plus important de l'établissement de groupes témoins est l'appariement le plus étroit possible des facteurs observés qui pourraient prédisposer une personne à réussir. Un bon exemple serait l'influence sur le revenu des études antérieures au programme : il y a une corrélation positive bien établie entre les études et le niveau de revenu (Lalonde, 1995). Idéalement, par conséquent, il faudrait que le groupe expérimental et

le groupe témoin soient similaires quant au niveau de scolarité antérieur au programme. Les évaluateurs devraient donc chercher un groupe témoin dont les moyennes et les répartitions, au chapitre des études, ressemblent le plus possible à l'échantillon des personnes qui suivent une formation.

Heckman et coll. (1995) soutiennent avoir élaboré une bonne méthode d'appariement fondée sur plusieurs variables (dans Heckman et Smith, 1995b). Pour établir des groupes témoins vraiment comparables, il est essentiel, affirment-ils, d'obtenir des renseignements sur les « transitions par rapport à l'activité », en particulier le passage de l'emploi au chômage, et de l'extérieur du marché du travail vers le chômage. Parmi d'autres variables d'appariement importantes, ces auteurs mentionnent l'âge, les études, la situation matrimoniale et le revenu familial.

La situation géographique et la période visée (p. ex., les deux groupes touchaient des prestations d'aide sociale pendant une période comparable) sont d'autres variables d'appariement importantes (voir Friedlander et Robins, 1995). Pour déterminer les variables à utiliser, il faut savoir quels sont les renseignements qui sont disponibles et connaître les buts du programme et de l'évaluation.

Il existe plusieurs méthodes différentes d'appariement : l'appariement selon l'admissibilité, l'appariement selon des cellules et l'appariement statistique sont des méthodes très répandues.

Dans l'appariement selon l'admissibilité, il s'agit de choisir les cas à partir d'un échantillon représentatif de la population qui répond aux exigences d'admissibilité du programme de formation. Mentionnons, par exemple, le fichier longitudinal de DRHC dans lequel on trouve le nom de tous les bénéficiaires d'a.-e. ainsi que les cours de formation qu'ils ont suivis. On pourrait choisir dans cette base de données à la fois ceux qui ont suivi une formation et ceux qui se sont abstenus (si la formation est axée sur les bénéficiaires de l'a.-e.).

Dans l'appariement selon des cellules, aussi appelé « appariement stratifié », les observations individuelles dans les deux échantillons sont réparties en cellules définies selon les caractéristiques qui prédisent la variable des résultats. On pourrait, par exemple, attribuer au préalable des membres de l'échantillon à des cellules fondées sur la région, sur la date de demande d'a.-e., sur l'âge, etc. Les cellules ne regroupant aucun individu qui suit une formation sont éliminées, celles qui ne contiennent que quelques cas sont combinées. Certaines études permettent ensuite la pondération des variables pour l'égalisation des répartitions entre les groupes. (Fraker et Maynard, 1987)

Pour construire un groupe témoin selon un appariement statistique (p. ex., par la méthode des plus proches voisins), il faut choisir un individu dans la population admissible pour chaque personne qui suit une formation, et ce, en fonction de la qualité de l'ajustement sur les points prévus sur les résultats prévus, ou de la qualité de l'ajustement sur les caractéristiques mises en corrélation avec le résultat (Fraker et Maynard, 1987). Les caractéristiques fondamentales de chaque personne qui suit une formation sont appariées avec celles de chaque personne qui n'en suit pas, et le dernier individu qui ressemble le plus à celui qui suit une formation est choisi.

Fraker et Maynard (1987) ont conclu que la population générale est une source insatisfaisante d'échantillons témoins. Friedlander et Robins (1995) ont indiqué que l'appariement statistique a produit des améliorations médiocres dans l'exactitude des estimations quasi expérimentales. Dickinson, Johnson et West (1987) ont signalé que le type de méthode d'appariement (appariement selon les cellules ou appariement statistique) n'a pas d'effet sur les résultats. Riddell (1991), dans son examen des évaluations de la formation, a convenu que la méthode d'appariement utilisée ne semble pas avoir eu des répercussions sur les estimations de l'effet même si les bases de données qui mettent l'accent sur les membres des groupes cibles peuvent accroître l'exactitude des estimations quasi expérimentales de l'effet du programme.

Par conséquent, vu que l'on peut de toute façon tenir compte de l'effet des caractéristiques individuelles dans le modèle, il est raisonnable de conclure que *le choix de la méthode d'appariement n'est pas aussi important que l'assurance que le groupe témoin satisferait aux exigences d'admissibilité du programme.*

9.0 Application des enseignements : le choix des groupes témoins pour les PEMS

Les enseignements tirés d'évaluations quasi expérimentales antérieures peuvent servir à l'établissement de groupes témoins régionaux pour les PEMS.

- À ce jour, il existe peu d'indications voulant que la méthode d'appariement utilisée influe beaucoup sur l'exactitude des estimations découlant des quasi expériences. Par conséquent, la voie la plus simple à suivre serait d'assurer que le groupe témoin réponde aux exigences d'admissibilité pour le programme de formation. *Le groupe témoin devrait tout au moins être tiré de la même population que le groupe expérimental.* Pour les programmes PEMS, lesquels visent principalement les bénéficiaires d'a.-e., on pourrait utiliser le fichier longitudinal de DRHC pour le choix des sujets des groupes témoins, par exemple. La base de données SNSE pourrait également être une source utile de personnes pour les groupes témoins.

- Pour chacun des volets des programmes PEMS, il existe des critères d'admissibilité précis qui doivent être pris en compte dans le choix d'un groupe témoin. Par exemple, les subventions salariales ciblées sont axées principalement sur les personnes qui font face à des obstacles particuliers à l'emploi, comme des déficiences. Le cadre échantillon pour le groupe témoin devrait provenir du sous-ensemble de la population de l'a.-e. qui se heurte également à ces obstacles.

- Dans le cas des programmes régionaux PEMS, il serait sage d'aller un peu plus loin que le simple appariement selon l'admissibilité : il semble manifeste que les groupes témoins devraient au moins provenir de la même région. Dans le cas du Fonds transitoire pour la création d'emplois, le groupe témoin devrait être choisi dans les mêmes milieux que le groupe expérimental (c.-à-d. ceux dont le taux de chômage est d'au moins 12 %). Si on le souhaite et pourvu que les renseignements nécessaires soient disponibles dans les bases

de données utilisées, il existe d'autres variables d'appariement qui peuvent être importantes, comme les changements d'activité, l'âge, les études, la situation matrimoniale, le revenu familial et la période.

- La stratégie d'échantillonnage dépend de la méthode d'appariement employée. La modèle la plus simple - et qui ne serait pas moins satisfaisante que d'autres plus complexes, selon la littérature - consisterait en premier lieu à limiter la population aux personnes dans la région qui bénéficiaient de l'a.-e. pendant la période où le programme était offert (plus tout autre facteur d'admissibilité propre à chaque volet des PEMS). On pourrait ensuite choisir un simple échantillon aléatoire. Il serait plus difficile mais pas impossible d'appliquer une stratégie d'appariement plus précise. Par exemple, si on voulait faire un appariement selon le sexe, l'âge et la scolarité, on pourrait établir une série de variables binaires (0-1) : hommes = 0, femmes = 1; moins de 30 ans = 0, plus de 30 ans = 1; études secondaires non terminées = 0, études secondaires terminées = 1. Par la suite, on pourrait ensuite calculer une variable à trois chiffres pour chaque participant et non-participant, où le premier chiffre représenterait le sexe, le deuxième, l'âge, le troisième, la scolarité (p. ex., une femme ayant un diplôme d'études secondaires et âgée 29 ans = 101). Finalement, on pourrait avoir recours au logiciel aléatoire pour déterminer un correspondant (ou plus d'un correspondant) pour chaque participant.

- Pour ce qui est de la taille de l'échantillon, on devrait avoir recours aux des modèles courants, en ayant à l'esprit le fait qu'elles calculent les marges d'erreur pour la taille de l'échantillon définitif plutôt que de l'échantillon original. Voir à l'annexe D la façon de choisir la taille d'échantillon qui convient.

10.0 Conclusion

Nous avons présenté dans ce document les éléments fondamentaux de la méthode d'évaluation quasi expérimentale, en mettant en relief des considérations sur la façon de tenir compte de l'effet du biais de sélection et sur le choix des groupes témoins. La méthode quasi expérimentale est habituellement la meilleure modèle à utiliser pour l'évaluation des programmes sociaux parce que le modèle comportant un seul groupe est insuffisante pour établir la relation de cause à l'effet, tandis que le modèle expérimentale avec deux groupes est souvent impossible à appliquer. Avec cette modèle, toutefois, il faut faire appel à des méthodes analytiques complexes pour isoler l'effet du programme par rapport à de nombreuses autres causes possibles.

Annexe A : Expériences véritables et quasi-expériences

On assiste, depuis une décennie, à un débat sur les qualités respectives de l'expérience véritable par rapport à la quasi expérience. Voici un exposé des arguments de chaque partie.

Le camp de l'expérimentation

Ceux qui favorisent le modèle expérimentale s'interrogent sur la validité de modèles non expérimentales très raffinées pour l'évaluation de l'effet des programmes de formation et d'emploi en raison de la difficulté apparente d'obtenir des estimations fiables de l'effet des programmes du marché du travail. Ils prétendent que les économistes qui font appel à des techniques quasi expérimentales n'ont pas vraiment réussi à isoler les effets des programmes (c.-à-d. à éliminer le « biais de sélection »)

Des universitaires tels Ashenfelter et Card (1985), Barnow (1987) et LaLonde et Maynard (1987) prétendent que les résultats de douzaines d'études économétriques ont été si variés qu'il pourrait n'exister aucune façon valable de bien mesurer les effets des programmes, sauf dans le cadre d'une évaluation expérimentale avec attribution aléatoire aux groupes de formation ou aux groupes témoins.

Les conclusions de Lalonde (1986) et de Lalonde et Maynard (1987) ont particulièrement démoli les arguments en faveur des modèles non expérimentales. Pour évaluer l'exactitude de telles modèles, ils ont comparé les résultats d'une expérience véritable du Projet national de démonstration d'emploi supervisé à ceux qui ont été obtenus par diverses méthodes non expérimentales largement utilisées, pour voir s'ils pourraient estimer avec exactitude les effets véritables du programme. Ils ont conclu qu'il ne semble pas exister de modèle quelconque [au moyen de méthodes non expérimentales] que les chercheurs pourraient utiliser en toute confiance pour reproduire les résultats expérimentaux du Programme d'emploi supervisé. En se fondant sur les mêmes faits, Fraker et Maynard (1987) ont conclu ce qui suit :

> Cette analyse a montré que les résultats peuvent être sérieusement faussés selon la population cible, le groupe témoin choisi et le modèle analytique utilisé. Plus important encore, il n'existe actuellement aucune façon de déterminer a priori si les résultats des groupes témoins procureront des indicateurs valables des effets du programme. (p. 216)

Des travaux plus récents sont venus confirmer les conclusions principales tirées de ces travaux, soit que les estimateurs quasi expérimentaux sont faussés et sont sensibles à des changements mineurs des caractéristiques du modèle. Friedlander et Robins (1995) ont évalué deux stratégies quasi expérimentales classiques en fonction de données expérimentales tirées de quatre expériences de réforme de l'aide sociale : ils ont comparé le groupe expérimental dans un lieu

à un groupe témoin dans un autre lieu, et comparé les résultats du groupe expérimental avec un groupe témoin antérieur au programme dans la même région. Ils ont conclu que les estimations non expérimentales étaient habituellement très différentes des estimations expérimentales, surtout pour les échantillons de comparaison tirés de diverses régions. Ils ont également étudié deux techniques statistiques pouvant améliorer l'exactitude des estimations quasi expérimentales : l'appariement statistique pour la formation de groupes témoins étroitement appariés, et des « tests de caractérisation » pour évaluer statistiquement le modèle économétrique employé pour déterminer si les estimations qui en découlent sont exactes[7] . Ni l'une ni l'autre stratégie n'ont sensiblement amélioré l'exactitude des estimations non expérimentales.

Greenberg et Wiseman (1992) ont aussi donné leur appui à la méthode expérimentale en déclarant que, après quinze ans d'expériences (p. ex., sur le soutien du revenu, pour la démonstration du programme d'emploi supervisé), ils ont pu constater que l'attribution aléatoire est une méthode méthodologiquement supérieure à l'évaluation de programme et démontrer que de telles études sont faisables. En outre, les travaux sur les évaluations de l'*Omnibus Budget Reconciliation Act* (par Manpower Demonstration Research Corporation) ont convaincu les décideurs de l'U.S. Department of Health and Human Services - qui financent habituellement l'évaluation des programmes de formation - que l'attribution aléatoire doit servir de fondement à l'évaluation de la réforme des programmes d'aide sociale. Les évaluations de MDRC ont été généralement jugées excellentes, ce qui a contribué à faire de l'attribution aléatoire la méthode de choix.

L'expérimentation présente un autre avantage, soit que les résultats sont compréhensibles et sont convaincants pour les décideurs (Burtless, 1995). Sans les réserves compliquées associées aux quasi expériences, les analystes peuvent présenter des constatations expérimentales simples, comme celle-ci : « Le programme a relevé le revenu annuel des participants de 1 000 $ ». Étant donné cette simplicité, il est plus probable que les décideurs utilisent les conclusions de l'évaluation. Ils ne s'empêtrent pas dans un débat scientifique interminable et souvent non décisif pour savoir si les conclusions d'une étude donnée sont statistiquement valables. Les élus sont plus sujets à agir sur des résultats qu'ils trouvent convaincants (Burtless, 1995, p. 67).

Enfin, un comité de la National Academy of Sciences a indiqué que les conditions ci-après sont nécessaires (mais non suffisantes) à une recherche de qualité : *l'utilisation de l'attribution aléatoire pour former un groupe*; la stabilité opérationnelle raisonnable du programme avant l'évaluation définitive; une bonne couverture de l'échantillon et de faibles taux d'attrition de l'échantillon; des mesures des résultats qui correspondent bien aux objectifs du programme, tant immédiats qu'à long terme; et une période de suivi pour que les effets du programme aient le temps de se manifester ou de disparaître. (Gueron et Pauly, 1991)

---

[7] Par exemple, on peut voir si le modèle prédit correctement qu'il n'y aura aucun écart des résultats entre les personnes qui suivent la formation et celles qui ne la suivent pas pendant la période avant le début du programme. Si le modèle constate des écarts statistiquement significatifs entre les groupes avant le programme, il faut l'abandonner puisqu'il a échoué le test de caractérisation.

Ces études ont notamment convaincu de nombreux évaluateurs que les estimateurs expérimentaux sont meilleurs que les estimateurs quasi expérimentaux (p. ex., Burtless 1995; Friedlander et Robins 1995).

Le camp quasi expérimental

D'autres, cependant, ont contesté l'opinion selon laquelle les modèles expérimentales sont supérieures aux modèles non expérimentales. Au premier plan de ce groupe se trouvent James Heckman et ses collègues (c.-à-d. Heckman, Hotz et Dabos, 1987; Heckman et Smith, 1995). Ils allèguent que, s'il existe un ensemble de données suffisamment riches et qu'on a recours à de bonnes techniques de modélisation économétrique, il est possible d'en arriver à des estimations fiables des effets.

Pour contester sérieusement le consensus de plus en plus répandu voulant que l'expérience est la meilleure voie à suivre, ils ont dû récuser la conclusion de Lalonde et de ses associés selon laquelle aucune méthode quasi expérimentale existante ne peut produire des estimations proches des estimations expérimentales non biaisées. Heckman et Hotz (1989) ont appliqué certains tests de caractérisation simples à des modèles fondés sur des données de l'expérience d'emploi supervisé et ont constaté que les modèles les plus inexacts pouvaient être rejetés, mais qu'il restait un sous-ensemble de modèles avec lesquels ils ont pu obtenir des estimations des effets semblables à celles des résultats expérimentaux. Heckman et Smith (1995) ont présenté des arguments convaincants qui démolissent dans une certaine mesure les conclusions de Lalonde : la taille des échantillons était trop restreinte et les données géographiques étaient insuffisantes pour que les membres des groupes témoins soient placés dans le même marché du travail local que les participants; des données antérieures au programme n'étaient disponibles que pour une seule année, de sorte qu'il était impossible d'avoir recours à des stratégies économétriques efficaces (et que les estimations pouvaient subir l'effet du « fléchissement d'Ashenfelter »[8]); les études n'ont pas eu recours à une diversité de tests de caractérisation courants; et les méthodes non expérimentales ont connu des progrès importants depuis que ces études ont été faites. Par exemple, Heckman et Smith (1995b) ont constaté que la dynamique de la main-d'œuvre (c.-à-d. les déplacements entre l'emploi et le chômage, ainsi que l'entrée dans la population active et la sortie de la population active), de même que d'autres facteurs connexes comme l'âge, la scolarité, la situation matrimoniale et le revenu familial, peuvent servir à la formation de groupes témoins qui sont « virtuellement identiques » aux groupes expérimentaux, de sorte qu'on peut tenir compte de l'effet du biais de sélection.

---

[8] Il s'agit du phénomène selon lequel le revenu des participants au programme de formation ont tendance à fléchir juste avant le début de la formation, parce que le chômage est souvent l'élément déclencheur pour suivre le cours. Par conséquent, les estimateurs de différences avant et après tendent à surestimer l'effet du programme. (Ashenfelter et Card, 1985)

Heckman et Smith (1995) soulèvent également des objections sérieuses, à la fois théoriques et empiriques, relativement à la méthode expérimentale :

- La randomisation peut modifier le bassin de personnes admissibles au programme ou changer le comportement des participants, ce qu'on appelle le « biais de randomisation ». Par exemple, pour former un groupe témoin, il peut arriver qu'on doive élargir le bassin de participants possibles, habituellement en élargissant certains critères d'admissibilité.

- S'il existe de proches substituts pour l'intervention expérimentale, l'hypothèse selon laquelle le groupe témoin ne subit aucune intervention est annulée dans la mesure où les membres profitent de la formation, ce qui entraîne un « biais de substitution ». Par exemple, certains programmes de formation consistent en l'achat de places dans des collèges communautaires; les non-participants peuvent choisir de suivre le même cours collégial s'ils paient eux-mêmes ou s'ils trouvent une autre source de financement.

- Les données expérimentales ne peuvent répondre à un grand nombre de questions qui intéressent au plus haut point les décideurs, notamment l'effet médian du programme et la proportion de participants subissant un effet positif (ou négatif) du programme. Il est impossible d'estimer des paramètres qui dépendent de la distribution combinée des résultats dans le groupe expérimental et le groupe témoin. C'est seulement si le problème d'évaluation est défini exclusivement en fonction des moyens qu'on peut dire que l'expérience procure une réponse précise. (Heckman et Smith, 1995, p. 22)

- L'évaluation expérimentale indique aux décideurs si un programme fonctionne ou non. Elle ne précise pas généralement pourquoi un programme fonctionne ou ne fonctionne pas.

- En raison de certains facteurs organisationnels, il est difficile de procéder à une attribution aléatoire dans certains cas et de faire le meilleur travail possible dans d'autres cas. Le personnel peut saboter le processus parce qu'il est opposé à l'attribution aléatoire. En raison des coûts, on peut décider de ne pas procéder à la sélection au point optimal dans le processus de décision[9] . La randomisation en plusieurs étapes, qui est coûteuse et qui perturbe le programme, peut être nécessaire à la production d'estimations des effets de services différents.

- Il faut souvent procéder à des analyses non expérimentales complémentaires pour remédier aux lacunes des modèles expérimentaux.

---

[9] La randomisation peut se produire à n'importe quelle étape du processus : le participant devient admissible, apprend l'existence du programme et son admissibilité, présente sa demande, est accepté, est évalué par le personnel, est affecté à certains services, commence à recevoir les services et termine le programme. La place optimale de la randomisation dépend des questions d'évaluation auxquelles on s'intéresse. Pour déterminer l'effet moyen, il faut procéder à l'attribution le plus tôt possible au début de la formation pour minimiser l'attrition.

Burtless (1995) convient que les évaluations expérimentales comportent de telles lacunes, mais il rétorque que les quasi expériences ont parfois aussi les mêmes (p. ex., les quasi expériences peuvent aussi être coûteuses et causer des perturbations); aussi, elles sont généralement affligées de problèmes statistiques plus graves que ceux qui se présentent dans les essais randomisés. En outre, ajoute-t-il, aucune règle propre à le modèle expérimentale n'interdit aux chercheurs de faire appel à des méthodes non expérimentales pour l'analyse des données.

Conclusion

Le jury en a encore pour longtemps à délibérer pour décider si les modèles quasi expérimentales peuvent bien tenir compte de l'effet du biais de sélection. Il est prudent de conclure que les modèles expérimentales sont supérieures sous cet aspect critique. Mais étant donné les nombreux problèmes qui leur sont associés, l'expérimentation est peu pratique pour un grand nombre, sinon la plupart des évaluations. La modèle quasi expérimentale est souvent la méthode la plus pratique à adopter pour l'évaluation des programmes de formation.

Annexe B : Brève représentation mathématique des quasi-expériences

Le problème de l'évaluation de l'effet d'un programme social dans un contexte non expérimental peut être représenté comme suit (Moffitt, 1991) :

$$Y^{**}_{it} = Y^{*}_{it} + \alpha$$
$$\alpha = Y^{**}_{it} - Y^{*}_{it}$$

où

Y^{*}_{it} = niveau de la variable du résultat pour la personne i au moment t si elle n'avait pas participé

Y^{**}_{it} = niveau de la variable du résultat pour la même personne au même moment si elle avait participé antérieurement

L'évaluation vise l'estimation de l'effet de l'intervention. C'est-à-dire que nous souhaitons estimer pour les personnes qui ont participé ce que Y aurait été si elles n'avaient pas participé. Manifestement, nous ne pouvons connaître Y^{*}_{it} puisque ces individus ont *effectivement* utilisé le programme. Nous remplaçons donc Y^{*}_{it} des non-participants pour estimer a :

$$\alpha = E(Y^{**}_{it}|d_i=1) - E(Y^{*}_{it}|d_i=0)$$

où $d_i = 1$ si la personne i a participé
$d_i = 0$ si la personne i n'a pas participé
et que | indique « à condition que » de sorte que le premier terme du côté droit de l'équation ci-dessus est la valeur moyenne de Y pour les participants, et le second terme du côté droit est la valeur moyenne de Y pour les non-participants.

En langage clair, nous estimons l'effet de l'intervention en estimant la valeur attendue (E) de Y, par exemple, le revenu annuel, pour les personnes qui ont participé à un programme de formation, et en soustrayant la valeur attendue de Y pour ceux qui n'ont pas participé. C'est seulement si $E(Y^{**}_{it})$ antérieurement au programme pour les participants est égal à $E(Y^{*}_{it})$ antérieurement au programme pour les non-participants qu'il n'y a pas de biais. Mais c'est rarement le cas en raison du biais de sélection.

Annexe C : Comment tenir compte de l'effet du biais de sélection

La façon sans doute la plus répandue de tenir compte du biais de sélection est le *modèle en deux étapes* de Heckman (1979). Dans la première étape, il s'agit de modéliser la sélection pour le programme. Cela prend habituellement la forme d'une équation simple qui expliquent la participation ou la non-participation au programme[10] :

$$P = \beta X + U$$

où P est une variable binaire (1 pour les participants et 0 pour les non-participants), X est un ensemble de tous les facteurs observés qui peut représenter la participation au programme (p. ex., l'âge, le sexe) et U est un terme d'erreur aléatoire qui est présumé être distribué normalement pour tenir compte de facteurs non observés qui influent sur la participation au programme. À partir de cette équation, l'inverse du rapport de Mill est calculé, et il est ensuite intégré à une équation de résultat de deuxième étape pour l'estimation de l'effet du programme (habituellement par la méthode des moindres carrés) :

$$Y = \beta X + \alpha P + \delta M + U$$

où Y est le résultat évalué, X est un vecteur de variables observées, P est l'élément fictif de participation et M est l'inverse du rapport de Mill. Si les hypothèses sur lesquelles repose le modèle sont exactes, la méthode selon Heckman entraîne l'élimination du biais de sélection (δ), ce qui donne une estimation non biaisée de l'effet du programme (α). (La mesure de l'effet du programme est le coefficient estimé de la variable indicatrice de la participation ou de la non-participation au programme.) Si cette équation était estimée selon la méthode ordinaire des moindres carrés, sans l'inclusion du terme de correction du biais de sélection, les estimations pourraient être biaisées. Toutefois, si le modèle est bien défini, l'ajout de la variable de la « correction du biais de sélection » élimine ce biais possible, de sorte qu'on obtient des estimations non biaisées de l'effet du programme.

---

[10] Dans la plupart des cas, l'équation est estimée comme s'il s'agissait d'un modèle selon des probits, ce qui convient si le terme aléatoire dans cette équation est normalement distribué. (Le probit est une mesure de la probabilité fondée sur les écarts par rapport à la moyenne d'une distribution de fréquence normale. Il est analogue à la régression multiple, mais avec une variable dépendante dichotomique.) Il existe toutefois (même si elles ne sont pas encore répandues) des méthodes en deux étapes pour les situations où l'hypothèse de termes aléatoires normalement distribués est peu susceptible d'être attestée.

Il existe un autre moyen très efficace de garder fixes les différences entre les groupes, soit la méthode des *différences à l'intérieur des différences*. On recueille des données longitudinales pour les principales mesures des résultats, p. ex., le revenu, l'utilisation de l'aide sociale. Pour tenir compte des différences dans les échantillons de participants et de non-participants, on a recours à un estimateur longitudinal de l'effet du programme; cet estimateur tient compte du niveau de la variable du résultat antérieurement et postérieurement au programme, par opposition aux estimateurs transversaux qui n'utilisent les données que pour les résultats postérieurs au programme. Cet estimateur est fondé sur le changement de la variable des résultats pour les participants, antérieurement au programme et postérieurement au programme, en tant qu'estimateur du changement qui se serait produit pour les participants en l'absence du programme. L'effet moyen estimatif du programme est alors la différence entre le changement antérieur au programme par opposition au changement postérieur au programme de la variable du résultat pour les participants et le changement antérieur au programme par opposition au changement postérieur au programme de la variable du résultat pour les non-participants. Il est ainsi possible de déterminer l'effet d'accroissement du programme en tenant compte de l'effet des biais de sélection entraînés par les différences individuelles non observées. On peut ensuite faire une analyse multivariée pour montrer comment la taille de l'estimation des différences à l'intérieur des différences de l'effet du programme varie en fonction de diverses caractéristiques des personnes et du programme.

Voici comment cela se présente sous la forme d'une équation (Moffitt, 1991) :

$$Y = E(Y^{**}_{it} - Y^{*}_{i,t-1} | d_i=1) - E(Y^{*}_{it} - Y^{*}_{i,t-1} | d_i=0)$$

où t = le point après l'intervention, t-1 = le point avant l'intervention et

$Y^{*}_{it} - Y^{*}_{i,t-1}$ = le changement dans Y^{*}_{it} de t-1 à t si l'intervention n'a pas eu lieu

$Y^{**}_{it} - Y^{*}_{i,t-1}$ = le changement dans Y^{*}_{it} de t-1 à t si l'intervention a eu lieu

Les méthodes de la variable instrumentale (VI) sont largement utilisées dans les cas où les estimations selon la méthode ordinaire des moindres carrés peuvent être biaisées en raison d'une corrélation entre une ou plusieurs des variables explicatives et le terme d'erreur aléatoire dans le modèle. Dans le contexte de l'évaluation et du biais de sélection, un tel biais possible existe en raison de la corrélation possible entre la variable de participation et de non-participation et le terme d'erreur aléatoire dans l'équation du résultat. On peut éliminer ce biais possible s'il existe une ou plusieurs « variables instrumentales » et qu'elles sont incluses dans le modèle.

Le modèle de base du résultat ou de l'effet du programme est le suivant :

$$Y = \beta X + \alpha P + U \quad (1)$$

Cette équation peut aussi s'exprimer ainsi :

$$Y = CW + U \quad (2)$$

où C = (ß α)' et W = (X P) en utilisant la notation matricielle.

L'estimateur des moindres carrés de (2) est exprimé ainsi :

c = (W'W)-1 W'Y où (W'W)-1 est l'inverse de la matrice (W'W). En général, cet estimateur est biaisé en raison de la corrélation entre W et U (c.-à-d. E{W'U} n'est pas égal à zéro, où E{ } représente l'opérateur des attentes).

L'estimateur de la VI de (2) est exprimé ainsi :

c* = (Z'W)-1 Z'Y où Z est la matrice des variables instrumentales. Cet estimateur est en général non biaisé parce que Z et U ne sont pas en corrélation, c.-à-d. que E{Z'U} est égal à zéro si Z est l'instrument approprié.

Annexe D : Détermination de la juste taille de l'échantillon

Quelle doit être la taille de l'échantillon pour une enquête donnée ? Tout dépend principalement de l'erreur tolérable, de la taille de la population, de l'importance de certains sous-groupes, du taux prévu de non-réponse et du budget disponible.

On entend par « erreur tolérable » la marge d'erreur pour l'enquête. À chaque fois que les résultats de sondages font les manchettes, la marge d'erreur – par exemple, plus ou moins 3 %, 19 fois sur 20 - est indiquée. La marge d'erreur indique au lecteur dans quelle mesure les conclusions du sondage sont exactes. Elle est fondée sur l' « erreur type », soit la mesure selon laquelle la moyenne de l'échantillon diffère de la moyenne de la population.

La marge d'erreur ajuste l'erreur type pour tenir compte des différences possibles entre l'échantillon et la population par le calcul de l' «intervalle de confiance » pour la moyenne de la population. On utilise généralement un intervalle de confiance de 95 % (c.-à-d. 19 fois sur 20). La marge d'erreur tolérable est habituellement de 3 % à 5 % (si elle doit être plus basse, le coût de l'enquête commence à monter rapidement).

La modèle traditionnelle pour les grandes populations est $n = 1{,}96^2\ p(1\text{-}p)/ET^2$, où n est la taille de l'échantillon à calculer, ET est l'erreur type tolérable et p est la proportion ayant la caractéristique qui est mesurée et (1-p) ont la proportion qui ne la possède pas (p. ex., si 48 % ont dit oui, 52 % doivent avoir dit non). Le chiffre de 1,96 représente le choix d'un intervalle de confiance à 95 % (dans une distribution normale, 95 % de la superficie sous la courbe est en deçà d'un écart type de 1,96 de la moyenne). Par exemple[11] , si une marge d'erreur de ± 3 %, 19 fois sur 20, est tolérable, il faut la taille d'échantillon suivante :

$$n = 1.96^2\ (.5*.5)/.03^2 = 1{,}068$$

La taille de la population est un facteur seulement lorsqu'elle est en deçà de 100 000 ou à peu près. En deçà, il faut utiliser, pour déterminer la taille de l'échantillon, un élément appelé « facteur de correction relative à une population finie ». Le facteur de correction est $(N - n / N-1)^{1/2}$, où N est la taille de la population et n est la taille de l'échantillon.

Si on intègre algébriquement ce facteur dans l'équation de la taille de l'échantillon, on a :

$$n = (1.96^2\ p(1\text{-}p)N)/\ (1.96^2\ p(1\text{-}p)) + (N\text{-}1)ET^2$$

---

[11] Par convention, p et 1-p sont établis au niveau le plus modéré — 0,5 pour chacun.

Par exemple, si l'évaluateur veut savoir combien de clients de l'a.e. il doit interroger dans un cadre échantillon de 2 500, avec un taux d'erreur de ± 3 % avec un intervalle de confiance de 95 % :

$$n = (1.96^2 * .25(2500))/ (1.96^2 * .25) + 2499(.0009) = 749$$

L'erreur d'échantillonnage associé aux sous-groupes est plus élevé que pour la totalité de l'échantillon, parce qu'il y a manifestement moins de cas. En règle générale, il devrait y avoir au moins 100 individus dans un grand sous-groupe qui sera analysé séparément. On obtient ainsi au moins une marge d'erreur de ± 10 % pour chaque grande strate, soit le maximum tolérable. (Rea et Parker, 1992)

À noter que, lorsqu'on choisit la taille d'un échantillon, il y aura toujours dans l'échantillon des personnes qu'on ne pourra trouver ou qui refuseront de collaborer. Il faut donc prévoir une marge pour les refus de répondre anticipés. Pour une taille d'échantillon *finale* de 1 000, avec un taux de réponse de 50 %, la taille d'échantillon *initiale* doit être de 2 000. Étant donné que le budget est fixe, il y a toujours un compromis entre la taille d'échantillon initiale et l'effort de réduire le taux de non-réponse. Il arrive trop souvent qu'on choisit une grande taille initiale et qu'on fasse trop peu d'effort pour réduire le taux de non-réponse, de sorte qu'il y a des répercussions sur l'erreur totale.

Bibliographie

Ashenfelter, O. and D. Card (1985) Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*. 67:648-660.

Barnow, B. (1987) The impact of CETA programs on earnings: a review of the literature. *Journal of Human Resources*. 22:157-93.

Burtless, G. (1995) The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives*. 9(2): 63-84.

Campbell, D.T. & J.C. Stanley (1971) *Experimental and Quasi Experimental Designs for Research*. Chicago: Rand McNally & Co.

Dickinson, K.P., Johnson, T.R., and R.W. West (1987) An analysis of the sensitivity of quasi experimental net impact estimates of CETA programs. *Evaluation Review*. 11:452-472.

Fraker, T. and R. Maynard (1987) The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*. 22:194-227.

Friedlander, D. & P.K. Robins (1995) Evaluating program evaluations: new evidence on commonly used nonexperimental methods. *American Economic Review*. 85(4):923-937.

Greenberg, D. & M. Wiseman (1992) What did the OBRA demonstrations do? In C.F. Manski & I. Garfinkel (Eds.) *Evaluating Welfare and Training Programs*. Cambridge: Harvard University Press.

Gueron, J.M. & E. Pauly (1991) *From Welfare to Work*. New York: Sage.

Heckman, J.J. (1979) Sample selection bias as specification error. *Econometrica* 47: 153-161.

Heckman, J.J., V.J. Hotz, and M. Dabos (1987) Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review*. 11:395-427.

Heckman, J.J.& V.J. Hotz (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *Journal of the American Statistical Association*. 84:862-877.

Heckman, J. and R. Robb (1985) Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, edited by J. Heckman and B. Singer, 156-246. Cambridge: Cambridge University Press.

Heckman, J.J & J.A. Smith (1995) Assessing the case for social experiments. *Journal of Economic Perspectives.* 9(2):85-110.

Heckman, J.J & J.A. Smith (1995b) Ashenfelter's dip and the determinants of participation in a social program: implications for simple program evaluation strategies. Unpublished manuscript, University of Chicago.

Heckman, J.J., H. Ichimura, J.A. Smith & P. Todd (1995) Nonparametric estimation of selection bias using experimental data. Unpublished manuscript, University of Chicago.

Lalonde, R.J. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review.* 76(4):604-620.

Lalonde, R.J. (1995) The promise of public sector-sponsored training programs. *Journal of Economic Perspectives.* 9(2):149-168.

Lalonde, R.J. and R. Maynard (1987) How precise are evaluations of employment and training programs: Evidence from a field experiment. *Evaluation Review*, II:428-451.

Mark, M.M. & T.D. Cook (1984) Design of randomized experiments and quasi experiments. In L. Ruttman (ed.) *Evaluation Research Methods.* Beverly Hills, CA.: Sage.

Moffit, R. (1991) Program evaluation with nonexperimental data. *Evaluation Review*, 15:291-314.

Riddell, C. (1991). Evaluation of manpower and training programmes: The North American experience. In OECD (Ed.) *Evaluating Labour Market and Social Programmes.* Paris:OECD.

Rubin, D.B. (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association.* 74:318-328.

Patton, M.Q. (1982) *Practical Evaluation.* Beverly Hills, CA.: Sage.

Posavac, E.J. & R.G. Carey (1980) *Program Evaluation: Methods and Case Studies.* Englewood Cliffs, N.J.: Prentice-Hall Inc.

Rea, L.M. & R.A. Parker (1992) *Designing and Conducting Survey Research.* San Francisco: Jossey-Bass.

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation: A Systematic Approach* (5th ed.). Newbury Park, California: Sage.

Stufflebeam, D.L. & A.J. Shinkfield (1985) *Systematic Evaluation.* Boston: Kluwer Academic Publishers.